MAPPING BICYCLIST ROUTE CHOICE USING SMARTPHONE BASED CROWDSOURCED DATA

A Dissertation
Presented to
The Academic Faculty

by

Aditi Misra

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Civil and Environmental Engineering

Georgia Institute of Technology
August 2016

MAPPING BICYCLIST ROUTE CHOICE USING SMARTPHONE BASED CROWDSOURCED DATA

Approved by:

Dr. Kari E. Watkins, Advisor
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Chris Le Dantec
School of Literature, Media and Communications
Georgia Institute of Technology

Dr. Patricia Mokhtarian
School of Civil and Environmental Engineering
Georgia Institute of Technology

Dr. Bistra Dilkina
College of Computing
Georgia Institute of Technology

Dr. Jorge Laval
School of Civil and Environmental Engineering
Georgia Institute of Technology

Date Approved: 23 June 2016

To Everyone Who Made This Journey Possible and Memorable

ACKNOWLEDGEMENTS

There are many people to whom I am indebted for their unconditional support, encouragement and guidance during my four years of doctoral study. First and foremost, I would like to thank my advisor Dr. Kari E. Watkins for her incredible positive energy and support that had never let me get intimidated by the challenges of this journey. She has been an awesome mentor, teacher and even a friend at times and this research would not have been possible without her active support, encouragement and advice.

I am also immensely thankful to Dr. Patricia Mokhtarian for her invaluable advice and guidance throughout this journey. Even within her busy schedule she had always found time to answer my questions and discuss my confusions, as well as to educate me in statistical modelling. From her I have learned to be positively critical and to be devoted to learning forever - a lot of who I am as a researcher today has been shaped by the time I was lucky enough to spend with her.

I also want to thank my committee members Dr. Bistra Dilkina, Dr. Jorge Laval and Dr. Chris LeDantec for their active interest in my research and progress. Their constructive suggestions, advice and guidance have contributed significantly in shaping my research goals and ideas over these four years. I would like to take this opportunity to thank Dr. Norman Garrick from University of Connecticut for inspiring my interest in transportation and for being a great mentor and support throughout my graduate studies. I would also like to thank Dr. Dipanjan Basu, my MS thesis advisor, for his kind support and guidance during the formative years of my graduate career – without his help this journey could be much more challenging.

My four years at Georgia Tech would not have been this memorable if I did not have Dr. Candace Brakewood, Dr. Susan Hotle, Dr. Adnan Sheikh, Dr. Aaron Greewood, Sarah Windmiller, Stephanie Amoning-Yankson, Janille Smith Collin, Alice Grossman, Simon Berrebi and Chieh Ross Wang by my side as friends, cohorts and even mentors at times – I cannot imagine how this journey would look like without the endless coffee chats, long stairwell meetings and critical tear down and reconstruction of each other's research ideas! I would also like to thank my friend Amrita Basak for her unwavering support in all my crazy endeavors, Ahmad Haider for keeping me fed during dissertation writing and my Art of Living meditation club members for keeping me sane through the last few months of dissertation writing.

No words are enough to express my gratitude towards my parents for letting me follow my dreams. I would like to take this opportunity to thank them for being so supportive, for passing me down their interests in all forms of learning and education, and, above all, for believing in me – I could never be here without their encouragement and sacrifices. I would also want to thank Jaydip for being immensely patient and understanding during this long journey – his faith and trust in my abilities have always motivated me to go that extra mile to make things work.

Last, but not the least, I want to thank Dr. Candace Brakewood and Dr. Aaron Goldman for always being there to keep me afloat during the rough times, and for teaching me, by their own examples, that no dream is big enough to be off limits and that nothing can stop you from achieving that dream if you believe in yourself. I hope and wish that in my future ventures, I will be able to spread this positivity among all people whose lives I touch.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

LIST OF FIGURES ix

LIST OF TABLES viii

SUMMARY x

CHAPTER 1. INTRODUCTION 1

- Background 1
- Motivation 7
- Research Overview 10
- Methodology Overview 14
- Contributions 23
- Dissertation Overview 26

CHAPTER 2. PARTICIPATORY PLANNING AND CROWDSOURCED DATA 30

- Introduction 30
- Crowdsourcing: Concepts, Platforms and Issues 32
- Crowdsourcing and Its Use In Transportation 40
- Conclusion 51

CHAPTER 3. SOCIO-DEMOGRAPHIC INFLUENCE ON RIDER TYPE CLASSIFICATION AND INFRASTRUCTURE PREFERENCE 53

- Introduction 53
- Literature Review 55
- New Cyclist Categories 62
- Methodology 64
- Discussion Of Results 95
- Limitations 97
- Conclusion And Future Research 98

CHAPTER 4. DATA CLEANING AND MAP MATCHING 101

- Introduction 101

Literature Review and State of the Practice 103
Methodology 109
Conclusion 125
CHAPTER 5. MODELLING CYCLISTS' CHOICE OF SHORTEST ROUTE AND DEVIATION FROM THE SHORTEST ROUTE 126
Introduction 126
Literature Review 129
Data 142
Multivariate Analysis 149
Model Results 161
Discussion 171
Conclusion 173
CHAPTER 6. CHOICE SET GENERATION 175
Introduction 175
Background and Motivation 177
Choice Set Generation Methodology 186
Results 191
Discussion 192
CHAPTER 7. ROUTE CHOICE MODELLING 195
Introduction 195
GPS Device Based Data Collection Efforts 196
GPS Enabled Smartphone Based Data Collection Effort 200
Methodology 203
Multivariate Analysis 205
Model Results 207
Discussion 214
Conclusion 215
CHAPTER 8. CONCLUDING REMARKS AND FUTURE DIRECTION 217
Future Research 218
Conclusion 220
REFERENCES 223

LIST OF TABLES

Table 1. Most Common Path Choice Models Used in Route Choice Modelling 23
Table 2. Basic Statistics for Socio-demographic Variables 68
Table 3a. Binary Logistic Regression Models 79
Table 3b. Ordinal Logistic Regression Models 80
Table 3c. Multinomial Logistic Models 81
Table 4. Odds Ratio for Multinomial and Ordinal Models with and without Cycling Frequency 82
Table 5. Means and Standard Deviations of Item Responses on Road Conditions and Facilities by Rider Type 89
Table 6. p – values for Pairwise t-test on Respondents' Ratings on Influence of Road Conditions and Facilities on Bicycling Propensity, Paired by Rider Type 90
Table 7. Exploratory Factor Analysis: Loadings 93
Table 8. Regression Analysis for Protected Environment, Route Impedance and Route Stress 95
Table 9. Infrastructure Related Variables Used in Models and their Modifications 148
Table 10. Trip Length as Function of Socio-demographic Characteristics 164
Table 11. Deviation from Network based Shortest Route as Function of Socio-Demographic Characteristics 168
Table 12. Segmented and Pooled Models for Gender 210
Table 13. Segmented and Pooled Models for Age 211
Table 14. Segmented and Pooled Models for Rider Type 211

LIST OF FIGURES

Figure 1: Cycle Atlanta User Interfaces: (a) Socio-demographic information, (b) Riding characteristics information, (c) Trip start screen, (d) Trip end and record screen, (e) &(f) Viewing trips, (g) Choosing a trip purpose, (h) Adding a note on anything about a particular spot 6
Figure 2. Cycle Atlanta Demographic Categories and Rider Characteristics 7
Figure 3. Path Overlap Among Possible Routes 20
Figure 4. Classifications of Crowdsourcing Systems (Doan et al. 2011, Steinfield et al. 2013, Erickson 2010) 36
Figure 5. Information Flow of the Transit Ambassador Program 48
Figure 6. Measure of Association between Variables 69
Figure 7. Socio-demographic and Riding Pattern Distribution of Cyclists across Rider Types 72
Figure 8. Socio-demographic Distributions of Pooled Survey Respondents across Rider Types 88
Figure 9. Original Uncleaned Data: (a) Raw GPS Points (b) Trip Lines Constructed from GPS points 111
Figure 10. Study Area for the Dissertation 116
Figure 11. Example of Matched Network Links (in red lines) to GPS Trip Points (black transparent Circles) 122
Figure 12. Cycle Atlanta Trips (a) Number of Trips Recorded by Users (b) Trip Distribution by Purpose (c) Trip Purpose Distribution across Age (d) Trip Purpose Distribution across Gender (e) Trip Purpose Distribution across Rider Type (f) Trip length Distribution (g) Trip Length across Age (h) Trip Length across Gender (i) Trip Length across Rider Type 152-155

SUMMARY

Bicycling has been identified as a critical component of livable communities, as it offers an environmentally friendly, cost-effective, congestion-reducing, and health-promoting mode of transportation for short trips. According to the National Household Travel Survey (NHTS 2009), nearly 40% of all personal trips in the U.S. are two miles or less, a reasonable bicycling distance. Although some of these trips may be secondary trips tied to the mode of primary trips (e.g. grocery shopping enroute to work or home), and are hence, non-convertible unless the primary commute mode is changed, the fact that only about 1.8 % of all those personal trips less than two miles are made on bicycles, is alarming. A major reason frequently cited for not adopting bicycling as a travel mode is a perceived lack of safety in facilities shared with high speed and volume traffic. To remedy the situation, ideally, all streets should be provided with separate bicycle facilities but agencies do not have enough funding nor enough right-of-way in many cases. Cyclists also differ widely in their perceptions of roadway safety and comfort and hence, possibly in their preference for infrastructure. To date, there are not enough data to understand cyclist preferences or how far the cyclists are willing to travel to access cycling facilities since bicyclists are a small and dispersed group and it is difficult to get data on their travel patterns through traditional traffic counts. Cycle Atlanta, a GPS based smartphone application (app) was developed at Georgia Tech in collaboration with the City of Atlanta to collect revealed preference route choice data of cyclists in Atlanta via crowdsourcing.

This research used the collected data to (1) validate the popular classification of cyclists into different rider types based on their comfort and confidence while also

modelling the underlying influence of socio-demographic attributes on self-perception of level of confidence and comfort; (2) develop a model to understand how far cyclists are willing to deviate from the shortest network distance based route and (3) design segmented route choice models for different types of cyclists based on their socio-demographics as well as their comfort and confidence level. In addition, a stated preference survey was analyzed to understand what factors influence the decision to use bicycle as mode of transportation. Finally, a data cleaning, curating, and map matching algorithm was developed as part of the research.

This research, one of the first studies to use crowdsourced data to analyze cyclist behavior, is unique in its focus on the influence of socio-demographic and attitudinal makeup of cyclists on their decision to bicycle and their route choice. The results from this research provide valuable insight for future planning and policy decisions. First, female and senior cyclists are found to be in general low confidence, low comfort riders and they significantly differ in their route choice and infrastructure preference from their more confident counterparts. Second, the assumption that with more riding experience cyclists become confident enough to share the street with vehicular traffic, is not without its caveats. Although cyclists with more riding experience tend to see themselves as more confident riders, preference for separate infrastructure pervades all rider types, as does the negative influence of high speed and volume traffic. Third, cyclists are generally found to shy away from longer trips and hence, when faced with the trade-off between a significant detour and safety concerns, they may not make the trip itself. Therefore, having a connected network close to the shortest distance path is important in encouraging newer and low confidence bicyclists. This research provides a model that

can be used to estimate acceptable deviation from any route based on road attributes and the cyclist characteristics.

CHAPTER 1

INTRODUCTION

Background

Traditionally, transportation planning in the U.S. has been automobile focused, resulting in marginalization of healthy and active modes of transportation like cycling and walking. Environmentally, this has led to increased air pollution; economically, this has made the country dependent on international fuel economy; and socially, this has brought about an alarming increase in obesity, heart disease and asthma among both adults and children (Sallis 2004). As a mode choice, bicycling can reduce overall congestion, air pollution and energy consumption while at the same time enable an active lifestyle and a low cost, equitable means of transportation. In view of all these, recently, the federal government has geared its policies towards promoting biking and walking, and several state and local transportation planning agencies have incorporated a bicycle planning module in their long term vision for the region. However, literature shows that although 40% of the trips made in U.S. are of bike-able distance, only 1.8% of such trips are bicycle trips (Pucher et al. 2011). Some of these trips may be secondary trips tied to the mode of the primary trips (e.g. picking up groceries en route to work or home), and are hence, non-convertible unless the primary commute mode is changed; nonetheless, the fact that less than 2% of all personal trips less than two miles are made on bicycles is alarming. This low subscription to bicycling has been generally attributed to safety issues (AASHTO 2012). Major factors contributing to negative safety perceptions are high speed limits, high traffic volumes, last mile disconnect in the network and absence of dedicated facilities for cyclists that can provide a physical separation from the vehicular traffic (Dill and Carr 2003, Buehler and Pucher 2011).

Studies reveal that a substantial increase in the number of bicyclists can be achieved by providing facilities for safe riding (Pucher and Buehler 2007) and therefore, it is important for the planning agencies to know where the cyclists prefer to bike and possibly their 'willingness to pay' for an added facility. Cities often try to organize the route network by balancing the connectivity of the network, shortest travel distances, parking locations, and traffic volumes (Dill 2004) – but this task is often difficult as perception of safety and comfort varies across level of experience of the cyclists, age, traffic characteristics and several other factors. Therefore, a better approach to understand cyclist route choice is to collect revealed preference data on the routes where the cyclists actually travel and then model the factors that influence the route choice decision of the cyclists.

Revealed preference data on cyclists are sparsely available for at least four reasons – first, bicycling trips often use by-lanes and short-cuts that are not manned during traditional traffic counts; second, bicycling trips also tend to happen during non-peak hours of commute, thus again not being counted during traffic counts; third, the automated counters are designed to detect vehicular metallic mass and therefore tend to underestimate bicycling trips; and fourth, since bicyclists constitute a marginal proportion of the total traffic, there are rarely separate count efforts employed for cycling trip counts. As a solution to such issues of data collection, GPS devices and GPS enabled smartphone applications have been developed to provide users the ability to record their trips by themselves. The earliest example of such an effort is the CycleTracks application developed at San Francisco County Transportation Authority (Hood et al. 2009) which has now been adopted by over a dozen cities across the U.S. This thesis is based off one such data collection effort fostered by the City of Atlanta, Atlanta Regional Commission and Georgia Department of Transportation in conjunction with Georgia Institute of Technology

where a GPS enabled smartphone application was built off CycleTracks but with substantial modifications to record cycling trip data in Atlanta.

Recently, the City of Atlanta started expanding its network of bicycle facilities to encourage people to bicycle more often. In doing so, they needed to understand the most travelled corridors, as well as particular streets that are avoided by cyclists even when those streets are the shortest connectors between any two points en route. For the purpose of data collection, collaboration was set up between the Georgia Institute of Technology and the City of Atlanta's planning office to develop a smartphone application that would help in collecting data from bicyclists who use these corridors and other city streets. The project was further facilitated by support from Atlanta Regional Commission who viewed the project as a means to foster "extensive public involvement by neighborhood residents, business owners, and the citywide cycling community" (The City of Atlanta, 2011).

The smartphone application created for this initiative was named Cycle Atlanta, after the name of the larger corridor planning project of which the app was a part, and was developed by an interdisciplinary team of researchers at the Georgia Institute of Technology. As mentioned, the application was based off of CycleTracks, although Cycle Atlanta was substantially updated to make better use of current features available in Apple Inc.'s proprietary mobile operating system (iOS) and Android as well as to include features that the City and local bicycle advocacy groups wanted in the application. The basic feature is trip recording, where the application uses the Global Positioning System (GPS) of the phone to record the location of the user once per second (Figure 1). At the end of the trip, the user is given the option to 'Save' the trip and only after the user saves the trip, the trip and related data are uploaded to a secure server. Once the trip is saved, the user can also specify the trip purpose and any related free-form note. Trip

purposes have been categorized as commute, school, work-related, exercise, social, shopping, errand, or other, enabling data users to segregate routes based on purpose as the infrastructure requirements and preferences may differ by the purpose of the trip. The free-form notes inform the city about the concerns of the users regarding particular routes and help in initiating correctional measures sooner.

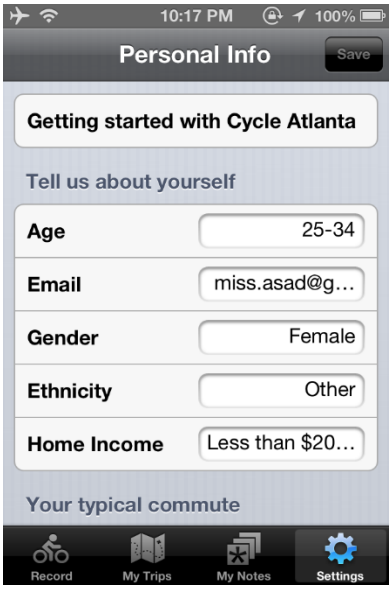


Fig. 1(a)

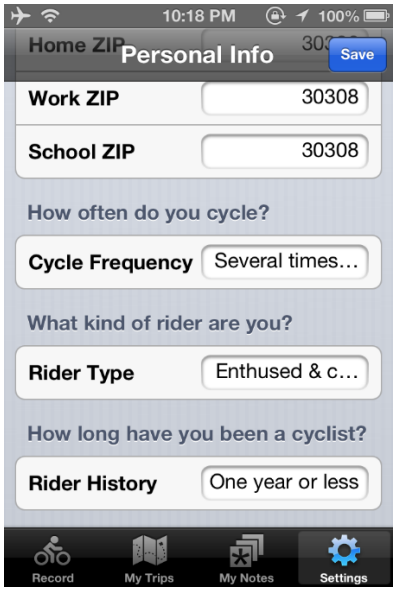


Fig. 1(b)

Figure 1: Cycle Atlanta User Interfaces: (a) Socio-demographic information, (b) Riding characteristics information, (c) Trip start screen, (d) Trip end and record screen, (e) &(f) Viewing trips, (g) Choosing a trip purpose, (h) Adding a note on anything about a particular spot

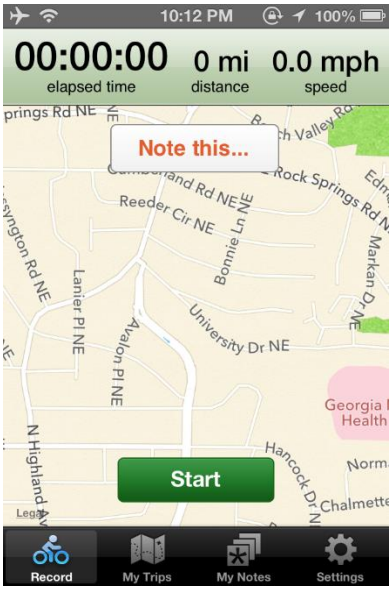


Fig. 1(c)

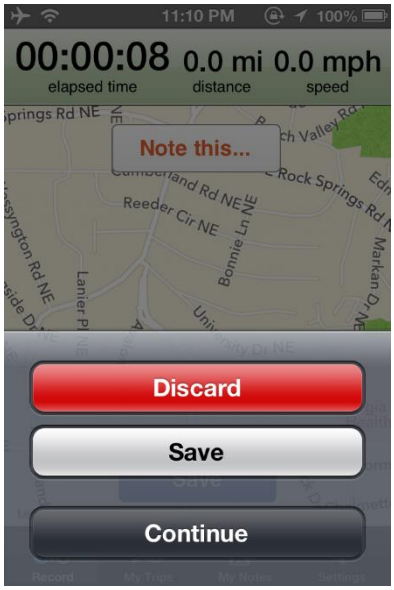


Fig. 1(d)

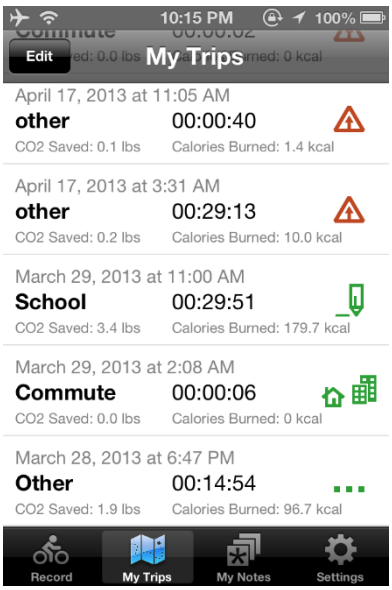


Fig. 1(e)

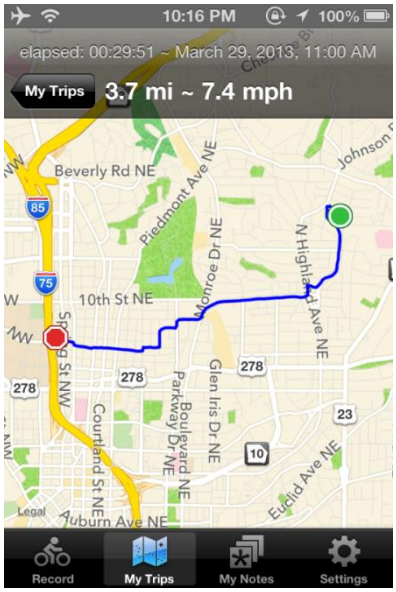
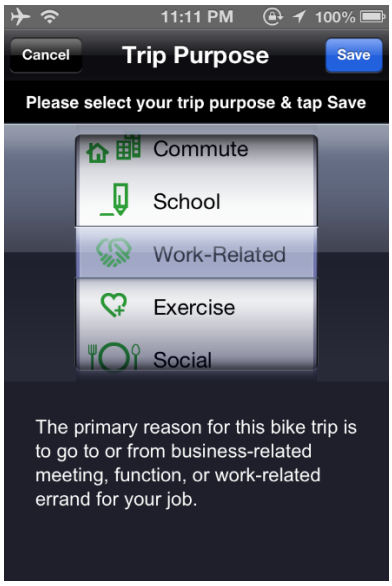


Fig. 1(f)

Figure 1: Cycle Atlanta User Interfaces continued



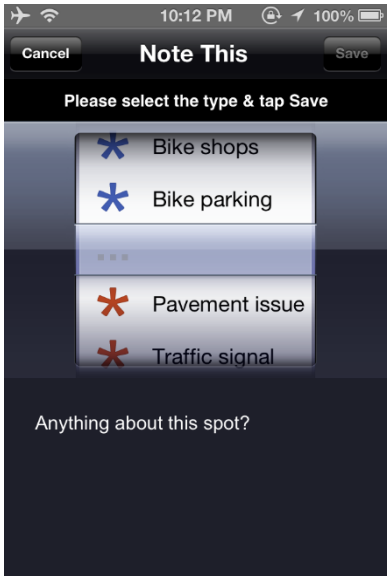


Fig. 1(g) Fig. 1(h)

Figure 1: Cycle Atlanta User Interfaces Continued

In addition to tracking cyclists' trips, the app also provides options to enter personal information, including age, email address, gender, ethnicity, home income, zip codes (home, work, and school), cycle frequency, rider type, and rider history. Figure 2 shows the group breakdown of the demographic and rider type categories in Cycle Atlanta. The breakdown of age, gender, income and ethnicity was kept similar to the breakdown as found in the household travel survey conducted by Atlanta Regional Commission while the rider type and rider history categories are exclusive and unique to the design of Cycle Atlanta. The rider history field allows users to specify how long they have been cycling and can choose from categories like ‘since childhood’, ‘several years’, ‘one year or less’ and ‘just trying it/just started’. This rider attribute can be used to see if people who biked from childhood are more likely to adapt to bicycling as a mode choice or if there is any relation between biking preferences and the years of experience the bicyclist has.

Age	Gender	Ethnicity	Home Income
Less Than 18	Male	White	Less than $20,000
18-24	Female	African American	$20,000-$39,999
25-34		Asian	$40,000-$59,999
35-44		Native American	$60,000-$74,999
45-54		Pacific Islander	$75,000-$99,999
55-64		Multi-racial	$100,000 or greater
65+		Hispanic/Mexican/Latino	
		Other	

Rider Type	Rider History	Cycle Frequency
Strong and fearless	Since childhood	Daily
Enthused and confident	Several years	Several times per week
Comfortable, but cautious	One year or less	Several times per month
Interested, but concerned	Just trying it/just started	Less than once a month

Figure 2. Cycle Atlanta Demographic Categories and Rider Characteristics

The other parameter, cyclist type, was adopted from a Portland State study (Dill and McNeil 2012) and modified to suit the needs of Cycle Atlanta. At its present form, there are four cyclist categories that the user can choose from – 'strong and fearless', 'enthused and confident', 'comfortable but cautious', and 'interested but concerned'. At its core, this new parameter actually represents an interaction term between the attitude of the rider and his/her comfort level with the city road network. This new parameter is thus best capable of presenting the varying need of different cyclist types and can actually be used as a proxy for risk aversion attitude for modeling bicyclist route choice.

Providing all such details is entirely optional on the part of the application user. Therefore, users are not required to enter information into any of the fields, and can still use the app to record trips if they choose not to share their personal information.

Motivation

The motivation of this research is derived from two main issues that it attempts to address. First, earlier studies that addressed cycling route choice problems, did so without any

consideration for attitudinal differences between cyclist types based on rider characteristics (for example, between 'strong and fearless' cyclists versus 'comfortable but cautious' cyclists), cycling experiences or socio-demographics like age or gender. It has been speculated that perception of safety varies across all these categories, but the effect of varying perception of safety on route choice decisions is a poorly studied area that needs further exploration. The Cycle Atlanta app collects optional demographic data from its users including age, gender, ethnicity, cycling frequency, rider type (comfort level), rider history (cycling experience), home income and zip codes (home, work and school). As will be evident from the literature review, the cycling experience and the type of cyclists are new features exclusive to the Cycle Atlanta dataset and inspire this research to identify how route choice preferences differ across rider type, age, gender, experience and any interaction thereof.

Second, planning agencies extensively use the cyclist classification proposed heuristically by Geller (2006) in connection with understanding perceived safety and comfort of cyclists with road infrastructure. However, the said classification is based on Geller's observation of the cyclists in Portland which already has a cycling culture – the classification was never tested for its generalizability to other cities, particularly with cities that are significantly different than Portland in cycling culture. In addition, the classification was never validated by user input – Dill and McNeil (2012) conducted a survey in Portland where they asked cyclists about their preferred cycling frequency and their comfort in using certain road types (e.g. main arterials with bicycle lanes vs without bicycle lanes) and then classified those cyclists into the categories proposed by Geller (2006) based on a combined metric calculated from the answers to the two questions mentioned before. This research aimed to determine the validity of the classification from the cyclists' perception instead of that of the analysts' by allowing users to self-classify

themselves into rider type categories with modifications to the categories to make it more inclusive of the safety concerned cycling enthusiasts that was felt missing in the original classification. The research is also motivated in establishing data interoperability between rider type and socio-demographics, so that agencies can use public domain socio-demographic data to understand any cyclist's comfort and confidence with different road types without the burden of additional data collection efforts.

On the data collection side, this research uses smartphone-application-based crowdsourced data for understanding the infrastructural need of the cyclists and their behavioral preferences. The advantages of using crowdsourced data is that it provides a less costly and labor intensive method of sampling for the planning agencies; while for the participants, it provides them with the flexibility of participation without any time and locational constraint. However, since the process is based on voluntary participation, there is a possibility of having self-selection bias in the collected data; i.e. any trend in the data may be heavily influenced by motivated and enthused cyclists rather than the infrequent and casual cyclists who are in more need of cycling infrastructure than the avid cyclists. The data may also suffer from systematic bias towards a young and high income generation, because of its dependence on technology for data input. Therefore, the ability of this research to realistically predict the behavioral preferences of the cyclists of Atlanta will add to the literature on the suitability and reliability of crowdsourced data for planning purposes.

In addition to these research goals, the aim of this research is to create a practical and useable tool that the planners and engineers can use to identify links that are critical for having a connected bicycling network between more frequently used origins and destinations. Research has shown that the decision to bicycle depends significantly on having a connected shortest route

between origin and destination. However, there is a threshold to the deviation from the shortest path that is acceptable for cyclists, and therefore it is important to understand how far a cyclist is willing to travel to access any new infrastructure created away from major arterials. This research is also motivated by the idea to contribute meaningfully in that direction by developing models to estimate propensity of a cyclist to deviate from the shortest route with a measure of how far that deviation is from the shortest route.

Research Overview

In accordance with the aims, this research was carried out in multiple related areas: (1) using the data collected via the smartphone application to understand the influence of socio-demographics of cyclists on their self – reported comfort and confidence as measured by the rider type label, (2) developing models for the cyclists of Atlanta to understand if and by how much cyclists deviate from the shortest route based on road attributes and cyclist characteristics (3) developing an open source data cleaning and map matching procedural standard, and finally, (4) developing route choice models for cyclists based on their socio-demographics and rider type classification and validating the hypothesis that infrastructure preference may differ significantly between different types of riders.

Socio-demographics and Rider Type

As mentioned before, Cycle Atlanta, a Geographical Positioning System (GPS) based smartphone application (app) was developed at Georgia Institute of Technology (Georgia Tech) in collaboration with the City of Atlanta to collect revealed preference route choice data of cyclists in Atlanta. Along with recording routes, the app provides the users the option to input demographics like age, gender, ethnicity and income, and rider characteristics like rider type, rider experience, and riding frequency while recording their trips. This research uses the data

collected through the Cycle Atlanta app to understand the route choice of the cyclists and how the choice is influenced by rider and route characteristics.

The first part of data analysis shows that socio-demographic variables and riding patterns are significant predictors of a cyclist's probability of self-classifying himself/herself into a particular category based on their comfort with presence or absence of cycling infrastructure and their interest in cycling. In particular, gender, rider history, and cycling frequency are significant in all the models. The results indicate that knowing a cyclist's demographic information can potentially help in classifying the cyclist into a particular rider type. In the future, this can help researchers to streamline surveys by replacing sociodemographic questions by a single rider type classification question. Alternatively, knowing the socio-demographics characteristics commonly available through census data and other surveys, researchers will also be able to predict the rider type and hence infrastructure preferences of people without having to undertake a new survey design for cyclists only. It will also help in understanding infrastructure and facility need of future cyclists who are not yet cycling and hence, there is no revealed preference data on their preference currently.

The results also direct attention to the requirement of segmented route and facility preference decision models for different cyclist types. Since the purpose of the route and facility preference analysis is to understand the requirements by rider types, segmented models based on rider type may enable a planner to better predict the choices of a future cyclist based solely on demographic information of the cyclists. Future route decision model research may therefore explore segmentation of the dataset to achieve better predictability.

From the second part of the analysis, it is evident that most route perception issues and facilities are viewed on a similar scale by cyclists as the mean scores on those facilities are quite

similar across rider types. Other results indicate that sociodemographic attributes and confidence levels influence infrastructure and facility preference. However, the model fits are substantially low indicating that rider level data are not sufficient to predict the route level decision process. Further investigation is necessary, as the literature shows that choice of route depends on route characteristics as well as rider characteristics like age and gender.

Data Cleaning and Map Matching

The next part of the study involved analyzing cycling trips as recorded by the users of Cycle Atlanta and mapping them onto the road network of Atlanta to retrieve characteristics of the links that the cyclist chose to travel on. This process proved to be one of the most time consuming parts of the project as (1) GPS recorded points are inherently noisy and often suffer from multitude of issues specific to the data collection platform thus calling for dedicated data cleaning and curating processes for the project, (2) snapping a series of GPS points to their corresponding correct road segments including intersections, requires a fully connected high resolution network of nodes and links which was not available initially; (3) cycling trips are unique in using short cuts and bylanes which are not part of the standard road network maintained by the agencies – but not having such short cuts in the map consistently makes the chosen route follow the link that the cyclist was actually avoiding by taking the shortcut; therefore, the road network file obtained from Atlanta Regional Commission had to be manually corrected to create connections and add short cuts and bylanes popularly used by cyclists, and finally, (4) available map matching algorithms were found to be heavily dependent on a nearest neighbor search or reconstruction of the route based on the shortest network based distance, none of which seemed suitable for a high resolution and spatially distributed network and cycling trips. A different algorithm was proposed that was expected to be computationally more efficient

than other approaches, but the approach still requires validation. Because of data and computational issues, the study area for this research was restricted to be within 5-mile radius of a chosen point in downtown Atlanta. Details of the data cleaning and the map matching processes used in this research is provided in Chapter 4.

Shortest Path and Route Choice Modelling

The final part of the research focused on developing shortest path and route choice models for cyclists in Atlanta. As mentioned already, not all major arterials have enough right-of-way to create separate bicycling facilities on the heavily used arterials. The main aim of this part of the research was to help agencies have an idea of what route attributes matter to cyclists the most when they choose a route and how far a cyclist might be willing to deviate from the shortest route to access a designated facility. Therefore, models were developed to understand (1) whether a cyclist would choose the shortest network-distance-based path for her trip based on both rider characteristics and road characteristics, (2) if a cyclist was not choosing the shortest network-distance-based path, how far was she willing to deviate from that path (3) if, and how the perception of routes (as determined by presence or absence of certain road attributes and infrastructure), and therefore, choice of a route may vary between different types of cyclists. The link attributes used for the analysis associated with the street network map were obtained from Atlanta Regional Commission and included traffic volume as measured by annual average daily traffic (AADT), posted speed limit, slope, presence of bicycle facility, presence of sidewalk and number of lanes. Percentage of truck traffic was initially considered but was found to be negligible along most corridors related to the study area. Factors like scenic beauty, parking and pavement conditions were not available in the current dataset – in future this study may be extended by using those variables in addition to the ones already used in order to have better

prediction power. Details of the models developed, background on related methodology and the results are provided in Chapters 5, 6 and 7.

Methodology Overview[1]

The central theme of this research is to understand how cyclists make decisions – whether to classify themselves into some ordered comfort and confidence category or whether to choose any route among an array of competing alternative routes. Human decision making is a complex process and there are multiple theories that have tried to explain it from different perspectives ranging from cognitive and behavioral theories to economic and logic based theories. The methodology used in this dissertation is based on discrete choice theory which originates from consumer economics and assumes that the attractiveness of a product or any alternative is evaluated by the decision maker as a function of its attributes. It also assumes that the said attractiveness can be mathematically expressed as a scaler valued function of the vector of attributes and a choice is made either by maximizing or minimizing that objective function (Ben-Akiva and Lerman 1985). Although the economic consumer theory allows for both discrete and continuous outcomes for the objective function, discrete choice methods, as the name suggests are concerned with discontinuous discrete outcomes for the objective function. In this regard, it is important to distinguish between compensatory and non-compensatory decision rules. In non –compensatory decision rules, choices are made by comparing only one attribute across all alternatives while for compensatory decision rules, choices are made based on a single

[1] This section is by no means a holistic coverage of the theory of discrete choice methods – that in itself is a wide and ever evolving area of research and is beyond the limits of this dissertation. Only those concepts, methods and models of discrete choice theory that are relevant and has been applied to answer policy questions relevant to this dissertation, are discussed here. Interested readers are directed to seminal texts by Ben-Akiva and Lerman (1985), Train (2001) as well as other textbooks on econometric analysis.

objective function composed of multiple attributes, none of which is singularly being maximized or minimized. For example, when comparing between three different modes of bus, train, and car, if a user makes a decision based on minimum travel time only, the decision rule used is non-compensatory. However, if the user is maximizing an objective function that simultaneously finds a trade-off between travel time, cost, and comfort, then the decision rule in use is called the compensatory decision rule and the objective function is called utility in consumer economic theory.

The discrete choice theory uses compensatory decision rule and assumes rationality of behavior on the part of the consumer i.e., if the person in the previous example prefers car as a mode of choice or commute, then under an identical situation, she will continue choosing that alternative. In addition, if she prefers car over bus and bus over train, then she will always prefer car over train, (consistent and transformative preferences). However, when put to experimentation, it was observed that people did not choose the same alternative every time even under identical situation and it was also observed that two people in identical situations faced with identical choices, make different decisions (Ben-Akiva and Lerman 1985). These findings led to a probabilistic definition of the discrete choice theory with two main approaches – the constant utility approach and the random utility approach. According to the constant utility theory, people do not choose the alternative that gives maximum utility – they choose according to a probability that is drawn from a distribution parameterized by the utilities. Since the focus of this dissertation is random utility, constant utility models are not discussed in detail here – interested readers are directed to Ben-Akiva and Lerman (1985) for detailed coverage on constant utility models.

Random Utility Models

The discrete choice models used in this dissertation are based on the random utility maximization (RUM) theory of economics. Random utility maximization theory assumes that people are rational and when faced with options, will choose the option that gives maximum utility. Formally, if a decision maker n has k alternatives available, then the decision maker n chooses the alternative i if the utility of alternative i U_{in} is greater than that of alternative j i.e., $U_{in} > U_{jn}$ for all i≠j. Formalized first by Manski (1977), the probability that a decision maker will choose alternative i over j given a choice set C_n is given by

$$P(i|C_n) = P[U_{in} > U_{jn}, all\ j \in C_n]$$

Utility U_{in} is a linear in parameter function that can be split into two components – (i) the part that is observable i.e., the attributes on which information is available (V_{in})and (ii) the part that is not observable by the modeler or the attributes that contribute to the choice decision but are not manifested to the modeler ε_{in}. Thus,

$$U_{in} = V_{in} + \varepsilon_{in} = \alpha_i \boldsymbol{X}_{in} + \beta_i \boldsymbol{Z}_{in} + \cdots . + \varepsilon_{in}$$

where $\boldsymbol{X}_{in}$ and $\boldsymbol{Z}_{in}$ are vectors of observable characteristics of the alternative and the decision maker

α_i and β_i are parameters to be estimated that give the maximum likelihood of choosing alternative i and ε_{in} is the unobserved part of the utility.

Based on this formulation of utility, the probability of choosing alternative i is given by

$$P(i|C_n) = P\left[U_{in} > U_{jn}, all\ j \in C_n\right]$$

$$= P\left[V_{in} + \varepsilon_{in} > V_{jn} + \varepsilon_{jn}\right]$$

$$= P[\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn}]$$

This expression is essentially a cumulative distribution for $\varepsilon_{jn} - \varepsilon_{in}$ such that

$$P(i|C_n) = P\left[\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn}\right] = \int I(\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn}) f(\varepsilon_n)\, d\varepsilon$$

where $I(.)$ is an indicator function that takes the value 1 when $(\varepsilon_{jn} - \varepsilon_{in} < V_{in} - V_{jn})$ is true and 0 when not. The distribution of the density $f(\varepsilon_n)$ gives rise to the different forms of discrete choice models of which the most commonly used one is the logit model with $f(\varepsilon_{ni})$ being Gumbel distributed with variance $\frac{\pi^2}{6}$ and having a form

$$f(\varepsilon_{ni}) = e^{-\varepsilon_{ni}} e^{-e^{-\varepsilon_{ni}}}$$

Since the difference between two Gumbel distributions is distributed logistic, $f(\varepsilon_n)$ is distributed logistic and hence the name of this class of models. The probability of choosing alternative i that case is given by $P(i|C_n) = \frac{e^{V_{in}}}{\sum_j e^{V_{jn}}}$ [2]

Model Estimation

The popularly used method of model estimation (i.e., estimating the values of α_i and β_i in the utility equation given above) is the method of maximum likelihood. From the utility equation, let us denote V_{in} as y and the vector of observable characteristics as $\boldsymbol{x}$. Let $\boldsymbol{\theta}$ be the vector of parameters $(i.e., \alpha_i, \beta_i s)$ partly or completely unknown that relate y and $\boldsymbol{x}$ such that $y \sim f(\boldsymbol{x}, \boldsymbol{\theta})$. With observations on samples ideally drawn at random from the population, a function is then created to estimate the unknown parameter vector $\boldsymbol{\theta}$. This estimate or estimator is given as $\widehat{\boldsymbol{\theta}}$. Maximum likelihood estimator is that $\widehat{\boldsymbol{\theta}}$ for which the observed behavior is most likely to happen.

---

[2] For details on derivation of the probability and on first and second order conditions of maximum likelihood, readers are referred to Train (2001) and Ben-Akiva and Lerman (1985)

For a sample of *N* observations, drawn at random from the population, the likelihood of observing the entire sample = product of the likelihood of observing each of them separately. If $\mathcal{L}^*$ is the likelihood function, then,

$\mathcal{L}^*(\theta_1, \theta_2, \theta_3, \ldots, \theta_n) = \prod_{n=1}^{N} \prod_{i \epsilon Cn} P_n\,(i)^{y_{in}}$ where $y_{in} = 1$ if observation n chose alternative i, and 0 otherwise. $P_n(i)$ for a linear in parameter logit model is given by $P_n(i) = \frac{e^{\boldsymbol{\theta' x_{in}}}}{\sum_{j \epsilon C_n} e^{\boldsymbol{\theta' x_{jn}}}}$

Since the product form is difficult to maximize, we generally maximize the logarithm of the likelihood function given as $\mathcal{L} = \sum_{n=1}^{N} \sum_{i \epsilon C_n} y_{in}\,(\boldsymbol{\theta' x_{in}} - ln \sum_{j \epsilon C_n} e^{\boldsymbol{\theta' x_{jn}}})$

Relevant Model Properties

The logit models rest on two very important assumptions – (i) the error terms are assumed to be independent and identically distributed, and (ii) the ratio of the probability of choosing one alternative to another only depends on those two alternatives and no other alternative in the choice set, also called the property of Independence of Irrelevant of Alternatives (IIA). The first assumption implies that the unobserved part of the utility for the alternatives should not be correlated across the alternatives i.e., there cannot be any attribute relevant to the choice decision common between alternatives that is not directly observed by the analyst. The second assumption IIA follows from the first assumption and although sounds intuitive and logical, it actually imposes the restriction that the alternatives cannot be related to each other in any way. If we go back to our original example of a decision maker choosing between car, bus, and train, and if for the sake of simplicity, we assume that the utilities of all the three options are the same, then, the probability of choosing car = probability of choosing bus = probability of choosing train = 1/3. Now if a fourth alternative is added of another bus serving an overlapping route with the bus already in the choice set, it is only natural and logical that this new bus will split the probability of choosing the original bus into half i.e. $P_{bus1} = P_{bus2} =$

$\frac{1}{6}$; $and\ P_{car} = P_{train} = 1/3$. However, logit models estimate the probabilities as $P_{bus1} = P_{bus2} = P_{car} = P_{train} = 1/4$ which underestimates the uncorrelated alternatives and overestimates the correlated alternatives.

This problem is particularly relevant in case of route choice model estimates where there are multiple overlapping links among the different alternatives being chosen. Consider for example the simple schematic shown in Figure 3. In travelling from origin O to destination D, a traveler has three alternative routes and let us consider that the utility of the path is only associated with its length. Since each of the paths have the same length T, they have the same utility and, then, under the assumption of IIA, the probability of choosing any of the three paths = 1/3. However, if path 1 and path 2 has overlap, then (1) part of the utility associated with path 1 and path 2 are exactly similar for that overlapping portion, including the unobserved errors and (2) it is more likely that users travelling along path 1 will explore path 2 rather than users from path 3 who are going in opposite direction, leading to path 1 and path 2 sharing the proportion of users originally on path 1 and path 3 retaining its own users. Let path 1 and path 2 be unique over the distance d but overlapping for the distance T-d. When d→T, therefore, the overlap is minimal, implying that all three routes are unique and the MNL estimate of the probability of choosing any path = 1/3 holds. However, as d → 0, path 1 and path 2 starts having more overlap and similar utilities implying that at d=0, probability of choosing either path 1 or path 2 is 0.5(1/3). For any intermediate values of d, the probability will vary between 1/3 and 1/6. This variation in probability due to overlapping utility is clearly not accounted for by logit models.

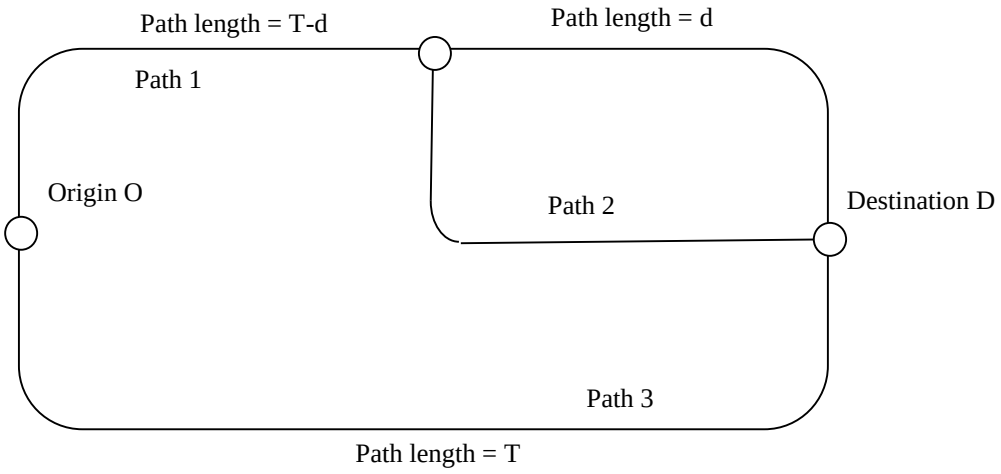


Figure 3. Path Overlap Among Possible Routes (Adopted from Ramming 2001)

Path Overlap Correction

As mentioned, major issue with route choice modeling in a large network with very similar alternative routes and overlapping links is the violation of the independence of irrelevant alternatives (IIA) property of logit structures. The issue warrants either (i) use of models with explicit inclusion of correlation terms like the Generalized Extreme Value (GEV) family or probit models which are computationally expensive and complicated or (ii) use of some correction factor along with the multinomial logit structure to account for the path overlap. Table 2 provides an overview of the different path overlap correction methodologies popularly used in route choice modelling problems. For a detailed discussion on all such methods, see Ramming (2001).

The first introduction of a correction factor in a logit structure is attributed to Cascetta et al. (1996). The model proposed by Cascetta et al. (1996), called C-logit, is based on MNL structure and involves estimating an additional term that is a similarity measure between a route and the other routes in a choice set. However, the most commonly used model specification is

the path size logit (PSL) presented by Ben-Akiva and Bierlaire (1999) and modified later by Ramming (2002). The probability of choosing path k according to the PSL structure is

$$P_k = \frac{\exp(V_k + \beta_{PS}.lnPS_k)}{\sum_{i\in C}\exp(V_l + \beta_{PS}.lnPS_l)}$$

where PS_k and PS_l are the path sizes of routes k and l respectively and β_{PS} is the parameter to be estimated. V_k and V_l are the observed utilities of path k and l respectively ($l \neq k$). C is the set of path choice alternatives of which i is any alternative. Path sizes are formulated differently by Ben-Akiva and Bierlaire (1999) and Ramming (2002) and are given respectively as

$PS_k = \sum_{a\in\Gamma_k}\frac{L_a}{L_k}\frac{1}{\sum_{l\in C}\delta_{al}}$ (Ben- Akiva and Bierlaire, 1999)

$PS_k = \sum_{a\in\Gamma_k}\frac{L_a}{L_k}\frac{1}{\sum_{l\in C}\left(\frac{L_k}{L_l}\right)^{\gamma_{PS}}\delta_{al}}$ (Ramming 2002)

where L_a is the length of link a, L_k is the length of path k, L_l is the length of path l, $\delta_{al} = 1$ if link a is traversed by alternative l and 0 otherwise, and γ_{PS} is a scale factor

While path size logit provides a computationally simple model, it only accounts for a part of the correlation and within MNL error structure. Frejinger (2007) estimated the PSL model with sampling correction while considering the full choice set of paths as the actual choice set, and results show that unbiased estimates are obtained only when the correction term is calculated on the full choice set.

Bovy et al. (2008) suggested a similar formulation of the problem but with modification of the path size term, and this model came to be known as Path Size Correction Logit (PSCL). The probability of choosing route k is given by

$$P_k = \frac{\exp(V_k + \beta_{PSC}.PSC_k)}{\sum_{l\in C}\exp(V_l + \beta_{PSC}.PSC_l)}$$

And the path size correction is given by:

$$PSC_k = -\sum_{a \in \Gamma_k} \frac{L_a}{L_k} ln \sum_{i \in C} \delta_{al}$$

where all terms have same interpretations as the previous equation (PSC_k and PSC_l are the path sizes of routes k and l respectively and β_{PSC} is the parameter to be estimated)

A completely different set of models not based on the logit formulation have been proposed to account for the correlation explicitly. These models are formulated based on the GEV structure and include Paired Combinatorial Logit (Prashkar and Bekhor 1998, 2000, Koelman and Wen 1998), Cross Nested Logit (Vovsha 1997, Prashkar and Bekhor 1998), and Generalized Nested Logit (Bekhor and Prashkar 2001, Wen and Koppelman 2001). While these models are good at accounting for link overlap issues, they are not used frequently because of the computational complexity and cost which far outweigh the benefits (Prato 2009).

Three different non-GEV and non-logit based models have been used – the multinomial probit (Daganzo and Sheffi 1977, Sheffi and Powell 1982), the Logit Kernel with Random Co-efficients or Mixed Logit model (Ben-Akiva and Bolduc 1996, McFadden and Train 2000, Jou 2001, Lam and Small 2001, Nielson et al. 2002, Nielson 2004), and Logit Kernel with Factor Analytic Approach (Bekhor et al. 2002, Frejinger and Bierlaire 2007). The computational costs for all of these models are significantly high, particularly for large networks which prevent their frequent use in route choice modelling problems. Therefore, for this research, the PSL model as proposed by Bovy (2009) is used.

Table 1. Most Common Path Choice Models Used in Route Choice Modelling

Name	Basic Structure	Modification	Formulation
C-Logit (Cascetta et al. 1996)	Multinomial Logistic	Additive Similarity Measure	$P_k = \frac{\exp(V_k+\beta_{CF}.CF_k)}{\sum_{l\epsilon C}\exp(V_l+\beta_{CF}.CF_l)}$
Path Size Logit (PSL) (Ben-Akiva and Bierlaire 1999; Ramming 2002)	Multinomial Logistic	Path Size Modification for Path Overlap	$P_k = \frac{\exp(V_k + \beta_{PS}.\ln PS_k)}{\sum_{i\epsilon C}\exp(V_l + \beta_{PS}.\ln PS_l)}$
Path Size Correction (Bovy et al. 2008)	Multinomial Logistic	Path Size Correction for Path Overlap	$P_k = \frac{\exp(V_k + \beta_{PSC}.PSC_k)}{\sum_{i\epsilon C}\exp(V_l + \beta_{PSC}.PSC_l)}$
Multinomial Probit (Daganzo and Sheffi 1977, Sheffi and Powell 1982), **Logit Kernel with Random Co-efficients/Mixed Logit model** (Ben-Akiva and Bolduc 1996, McFadden and Train 2000, Lam and Small 2001, Nielson 2002, 2004) **Logit Kernel with Factor Analytic Approach** (Bekhor et al. 2002, Frejinger and Bierlaire 2007)	Non GEV/Non Multinomial Logistic	Inherently do not have the IIA property	$P_{nk}(\beta_n) = \frac{\exp(\beta_n' X_{nk})}{\sum_{i\epsilon C}\exp(\beta_n' X_{nl})}$ $P_{nk}(\beta_n) = \int \frac{\exp(\beta_n' X_{nk})}{\sum_{i\epsilon C}\exp(\beta_n' X_{nl})} f(\beta)d\beta$

Contributions

This research is one of the first studies to use smartphone based crowdsourced data to understand and model the route choice of cyclists. The major contributions of this dissertation are:

(a) Developing an overarching classification framework for crowdsourcing systems based on existing multi-disciplinary literature on crowdsourcing and situating the current research in its context, thus setting a precedent for future transportation related crowdsourcing projects

(b) Examining the validity of a popular cyclist classification system that heuristically classifies cyclists based on their comfort and confidence with different infrastructure. This dissertation explores the relation between socio-demographic makeup of a cyclist and her self-classification in the categories of the above mentioned system. Since this classification system is presumed to be related to the infrastructure preference of the cyclists, the findings from this research can help in planning for preferred infrastructure for cyclists using publicly available socio-demographic data without any additional survey burden.

(c) Developing a model to understand a cyclist's propensity to choose the shortest network distance based path and how far from that path she is willing to deviate to access a route that she perceives to be preferable. None of the studies on bicyclist route choice have previously developed a model that can help the agencies to estimate the cyclists' willing ness to deviate from the shortest route based on route and rider characteristics. Therefore, this study makes a unique contribution in that area which is also hugely useful for practitioners. In the future, planners can use this model to estimate the optimum distance from the major arterial where they can create separate infrastructure for cyclists that will be effectively used by the cyclists.

(d) Developing route choice models for different types of cyclists to understand if route preferences of cyclists vary by their socio-demographic attributes or comfort and confidence. Since most cyclist data are heavily biased towards male and younger riders, any model estimate on a composite dataset tends to highlight preferences and willingness of such riders. Previous research hypothesized based on stated preference surveys that male and female cyclists as well as younger and older riders vary significantly in their perception of safety which can easily translate into their route choice decisions. However, bicyclist route choice models have always been estimated on pooled data, mainly because of the small size of the datasets and therefore, this hypothesis has never been validated via data. Cycle Atlanta, with its strong user base, provides the unique opportunity to explore and compare stated preference of the cyclists to their actual behavior and these models are designed for that purpose. The results from this dissertation will contribute in understanding if all cyclist preferences are similar or if there is reason to make special provisions for some particular group of cyclists. This study is also the first of its kind as none of the previous cyclist route choice models have looked into segmented models.

In addition, this dissertation also developed a data cleaning, curating and map matching algorithm that was used for this research and can be used for future GPS based data collection efforts. This dissertation also used a deterministic calibrated utility based choice set generation method that is computationally simpler than stochastic methods but gives similar coverage. Although the method is new in its approach, it is yet to be validated for computational speed and efficiency against existing methods before its potential contribution can be asserted.

In summary, this research will improve opportunities for bicycling, and therefore community livability, in urban and suburban areas by evaluating the bicycle network and

identifying the routes/ links that will have the most impact on cycling ridership should they be improved. This research is innovative as the analysis is based on crowdsourced observed and perception data collected in real time from actual cyclists, a new form of data collection. Furthermore, efforts to improve the bicycle route network using crowdsourced data is a powerful means of incorporating citizens in infrastructure improvement decisions, which will improve livability by maximizing the benefit of the bicycle infrastructure funding and empowering citizens to be more active in transportation decisions. In addition, this research provides a tool to assess how far a cyclist may deviate from the shortest network distance based route to access a route that she perceives as safe. From a policy and implementation perspective, this research will help planning agencies understand how changes in road attributes like traffic volume and traffic speed can help bicyclists share streets with vehicular traffic and use the shortest route to their destination. This research will also help in optimizing placement of cycling infrastructure away from the main arterial when there are issues with right of way by estimating how far the cyclists are deviating from the shortest path to access a facility, when controlling for everything else.

Caveats

The innovative use of a smartphone-based application to collect revealed preference cyclist route choice data has its own caveats. The Cycle Atlanta data suffer from issues of self-selection bias as it is a crowd sourced data collection system and people who participate are those who are sufficiently interested in the project, willing to share data, and willing to invest time without any personal gain. The Cycle Atlanta dataset is heavily dominated by male white cyclists in the age group between 25-44 years. This is not a representative sample of the population of Atlanta where 50% of the population is female and about 54% is African American (ACS 2012). However, there is currently no reliable estimate on the makeup of the cycling population of

Atlanta (Poznanski 2013), and hence it is difficult to comment on the representativeness of the Cycle Atlanta data with regard to the cycling population of Atlanta. We compared the sociodemographic distribution of the Cycle Atlanta users to the participants of the Atlanta Regional Commission cyclist survey and found no statistically significant difference. While this may mean that the cycling population of Atlanta is fairly homogenous, it may also be due to the advertisement of Atlanta Regional Commission's survey through pro bicycle channels that typically reach out to people with similar attitudinal preferences as the Cycle Atlanta users. It is an ongoing future research debate as to whether weighting the data by population proportion should be considered in the case of cycling data, as that may interfere with the representativeness of the collected data. A comparison with the U.S Census data (ACS 2014) on cyclists however, point to a 28% female and 72% male gender break up among cyclists nationwide which is similar to the break up that we see for the Cycle Atlanta data. Likewise, while the average cyclist across the nation is more likely to be younger, in the age group of 16-29, Cycle Atlanta data also shows a higher representation of younger cyclists, particularly in the age group between 25-44. Ethnicity-wise, African Americans tend to have lowest cycling rate across the nation while Hispanics tend to have the highest cycling rate (https://www.census.gov/prod/2014pubs/acs-25.pdf). However, in Cycle Atlanta, the dominant group is White while Black Americans are the second largest. This may be due to the population characteristics of Atlanta itself, which is heavily weighted by African Americans. The other point of departure for Cycle Atlanta from the national data is the income of cyclists – across the nation cyclists tend to be low income people (less than $10,000 annually) whereas Cycle Atlanta has a higher representation of the high income group (from $60,000 to greater than $100,000).

The other issue associated with the data is that by design, the data systematically misses the cyclists who do not own a smartphone. A study on smartphone ownership (Windmiller et al. 2014) has shown that this systemic bias affects people in older age groups, certain ethnic groups, and sometimes people in lower income groups. The Cycle Atlanta dataset is sparse in all of these categories – the users are mostly White and high income in the age group of 25- 44. It is difficult to estimate how much of the sparsity is caused by use of smartphone for data collection and how much is due to characteristics that define cyclists of Atlanta. However, the number of non-smartphone owning Atlanta Regional Commission participants and Cycle Atlanta participants were similar in number, therefore making up a similar portion of the sample. The models were run for the non-smartphone owners as a separate group and no statistically significant difference in infrastructure preference was noticed. Therefore, in spite of the biases in the data collected via smartphone app, any infrastructure requirement predicted based on Cycle Atlanta data can be assumed to hold true for non-smartphone users as well.

Dissertation Overview

The remaining chapters of the dissertation are arranged as follows:

Chapter 2 deals with crowdsourcing as a system of data collection – it provides a literature review of the existing crowdsourcing systems, develops a classification for such systems and situates the current research in its context.

Chapter 3 focuses on understanding if and how the self-perception of confidence and comfort as measured by the attribute ‘rider type’ varies across different sociodemographic segments like male vs female riders or younger vs older riders. This chapter also provides the analysis and results of a stated preference survey conducted by Cycle Atlanta team and Atlanta

Regional Commission which aims to understand how different road attributes can influence the propensity to bicycle. A literature review on cyclist classification and of stated preference studies on cyclists' infrastructure preference is also provided in this chapter.

Data cleaning, curating and map matching procedures developed and used for this research are presented in Chapter 4. This chapter also provides the relevant literature review on the current state of the art in GPS data cleaning and map matching efforts.

Chapter 5 models the willingness of the cyclists to choose or deviate from the shortest network distance based route and also by how far they are willing to deviate. Since the methodology used in this chapter are slightly different than that of the other chapters, a detailed overview of the methodology is provided within the chapter along with a literature review of previous related studies on cyclists' choice of the shortest path.

Chapter 6 and Chapter 7 focus on route choice modelling for cyclists in Atlanta. Chapter 6 deals with choice set generation or generation of alternatives for the chosen route while Chapter 7 provides the actual route choice models. Overview of existing choice set generation methodologies is presented in Chapter 6 while a literature review of existing bicyclist route choice models is presented in Chapter 7 as the methodological background relevant to that chapter has been covered in Chapter 1.

CHAPTER 2

PARTICIPATORY PLANNING AND CROWDSOURCED DATA

Introduction

Researchers have long emphasized the importance of public participation in the planning process as a critical component to the successful implementation of any plan *(*Innes 1998, Burby 2003, Slotterback 2010*)*. Broad public participation leads to "greater legitimization and acceptance of public decisions, greater transparency, and efficiency in public expenditures, and greater citizens' satisfaction" (Burby 2010). According to Burby, inclusion of stakeholders with varied interests and different backgrounds makes a plan comprehensive, acceptable, and more easily implementable (Burby 2010). Moreover, a participatory planning process effectively recognizes that "society is pluralist and there are legitimate conflicts of interest that have to be addressed by the application of consensus building methods" (Hague et al. 2003). With these traits in mind, participatory planning has the potential to involve broader and more diverse groups of people into a planning dialogue and hence, can bring in newer perspectives and ideas to the planning problem at hand (Rabinowitz 2013).

Recent research, however, suggests that citizen involvement at different stages and levels of planning is steadily declining in the U.S. (Skocpol and Fiorina 1999, Galston 2004, Pew Research Center 2013). This seems counterintuitive given the fact that over the last few decades, information accessibility and remote participation has been facilitated and made easier through the ubiquitous use of the internet and web-based social media. A wealth of emerging technologies has brought about significant new forms of communication and interaction, providing diverse new ways of documenting, sharing, and reflecting on the world at a truly global scale.

One possible reason for this decrease in citizen involvement may be that planners and policy makers have yet to embrace technology-mediated forms of participation and instead still rely on methods that require the physical presence of the participant. These methods limit the availability of the planning process by placing time and location constraints on participation and may also alienate or further disadvantage citizens for whom travelling to a planning meeting is neither physically nor financially viable.

One strategy for overcoming limited participation from interested stakeholders is to implement multiple methods of participation that participants can choose from, depending on their level of comfort and accessibility (Wagner 2012). Slotterback (2010) proposed that along with the traditional methods of public hearings and open-house meetings, more accessible modes of communication like project websites, web-based meetings and discussions may be adopted as a means of increasing public participation in the planning process. Toward that end, the purpose of this paper is to encourage the use of crowdsourcing platforms as a possible means of involving people from diverse walks of life to effectively participate in planning for transportation systems without putting additional financial burden on the transportation agency. This chapter highlights the successful use of crowdsourcing in a few transportation projects, providing examples of projects that have overcome many of the initial challenges of adopting crowdsourcing in transportation planning and establishing a robust starting point for future work.

This chapter is organized as follows: first, the concept of crowdsourcing is discussed along with a commentary on the existing platforms and types of crowdsourcing and the issues associated with crowdsourcing in general. Then, the crowdsourcing case studies in transportation planning are presented with reference to the different genres of crowdsourcing. The first group of case studies focuses on receiving feedback from users while the second group

focuses on use of crowdsourcing for data collection. A standalone example is provided at the end of the case studies sub-section as it deserves special mention because of its use of data quality editors to ensure data usability and validity, thereby addressing one of the biggest issues of crowdsourced data collection.

Crowdsourcing: Concepts, Platforms and Issues

At its conception, social computing focused mainly on building a network of collaborators and facilitating online communication between groups. This has eventually given rise to open source platforms and forums where people with similar motivation and outlook can come together to solve issues and to find answers to problems that affect their community. Crowdsourcing is one such example where an organizer or an organization is able to use the network of collaborators to solve a problem that would otherwise be cost or labor intensive or for which the available expertise within a defined organization is unavailable or insufficient.

Crowdsourcing has been alternately defined as: the outsourcing of a job (typically performed by a designated agent) to a large undefined group in the form of an open call (Howe 2006); a process that "enlists a crowd of humans to help solve a problem defined by the system owners" (Doan et al. 2011); or "a sourcing model in which organizations use predominantly advanced Internet technologies to harness the efforts of a virtual crowd to perform specific organizational tasks*"* (Saxton et al. 2013). Common across these alternate definitions is the notion that crowdsourcing invites all interested people to form an open forum of ideas that can eventually lead to a solution of the assigned problem. As Howe (2006) states, crowdsourcing utilizes the "latent potential of crowd" to achieve a solution to a problem that the crowd can relate to.

According to Saxton et al., crowdsourcing systems are characterized by three main features – the process of outsourcing the problem, the crowd, and a web-based platform for collaboration (Saxton et al. 2013). Outsourcing a problem generally implies getting a task done by outside sources even when it could have been performed by people within a system; in crowdsourcing, outsourcing is done in cases where either the in-house expertise has failed to produce a solution, or is an expensive means to produce a solution, or there is no in-house expertise available to use for solving the issue. Crowdsourcing systems also rely largely on an anonymous unidentified group of people ("the crowd") to come together willingly instead of the business sub-contract model of outsourcing where the task is performed by a previously identified and designated group of people or a company (Saxton 2013).

An important subset of the general crowdsourcing idea is the concept of citizen science, in which amateurs contribute to research projects in conjunction with the professional scientists. Goodchild used the term "citizen science" in describing crowdsourced geo-mapping, referring to the fact that information generated through crowdsourcing, although not of the level of a professional, helps in expanding the reach of science (Goodchild 2008). The nature of participation of the people in citizen science projects takes different forms depending on the type of the project and can range from data collection to data analysis, from instrument building to taking part in scientific expeditions. Recent citizen science projects tend to focus on utilizing the ever increasing reach and availability of electronic gadgets, particularly mobile phones and sensors, for data collection and monitoring purposes. In their experiments, Kuznetsov and Paulos (2010) and Kuznetsov et al. (2011) provided citizen scientists with sensors to monitor air and environmental quality, while the CycleTrack project in San Francisco used GPS enabled mobile devices to record cyclist trip data (Hood et al. 2011). Citizen science projects are gaining

popularity as an alternative to cost intensive data collection efforts, particularly in cases where the information needed is global in character, and are thus being increasingly used for planning and monitoring purposes.

Existing Crowdsourcing Platforms and Systems

Despite the advantages discussed in the previous section, crowdsourcing can only be successful if a platform exists that can provide open access to incorporate, modify, and synthesize data. There are four different versions of this shared platform – the wiki system, open source software, geocrowd mapping, and mash-ups using crowdsoucring data (Kitchin and Dodge 2011). Wiki systems are mainly centered on authoring information; open source software provides a platform to share and co-develop program source code; geocrowd mapping entails collecting, cleaning, and uploading GPS data; and mash-ups are combinations of some or all of these. While maintaining coordination among people coming from different backgrounds and motivations is a significant challenge, this voluntary coming together of a mass of people for a purpose is particularly useful in tackling problems that are large scale, e.g., mapping of a country.

Beyond the fundamental concept of providing an open access and participatory platform for a large group of people, crowdsourcing projects can be markedly different depending on the purpose of the project, the nature of involvement required, or if some special expertise is required for participation. Figure 4 schematically represents the different categorizations of crowdsourcing systems which are further discussed herein. Based on the nature of involvement of the participants in solving the problem, Doan et al. (2011) classified crowdsourcing systems as either explicit or implicit systems (Figure 4). Explicit systems are standalone systems where users participate and collaborate in executing a stated problem like answering questions via the

web, testing software and writing web content (e.g., Wikipedia). Within explicit systems there are four different types of tasks that users generally perform: (i) evaluating (e.g., book review), (iii) sharing (e.g., feedback on system performance), (iv) building artifacts (e.g., designing T-shirts at Threadless.com), and (v) executing tasks (e.g., collaborating on finding gold mining spots). Implicit systems can be standalone or piggyback depending on projects. In standalone implicit crowdsourcing systems, the system owners benefit from the indirect input provided by the users; the direct user input is used to solve a problem that is related to but not the same as the issue that the users of the system respond to. For example, although humans are more efficient at image recognition than computers, they are not necessarily willing to perform this task unless it is packaged in a form that attracts them. In the Extra Sensory Perception (ESP) game, the participants are shown images and asked to guess common words to describe those images as part of playing the game. Those words are then used to label the image (Doan et al. 2011). In piggyback crowdsourcing systems, the traces of the users are collected from an entirely different system – ad keywords generated based on Google and Yahoo search traces are examples of piggyback implicit crowdsourcing systems.

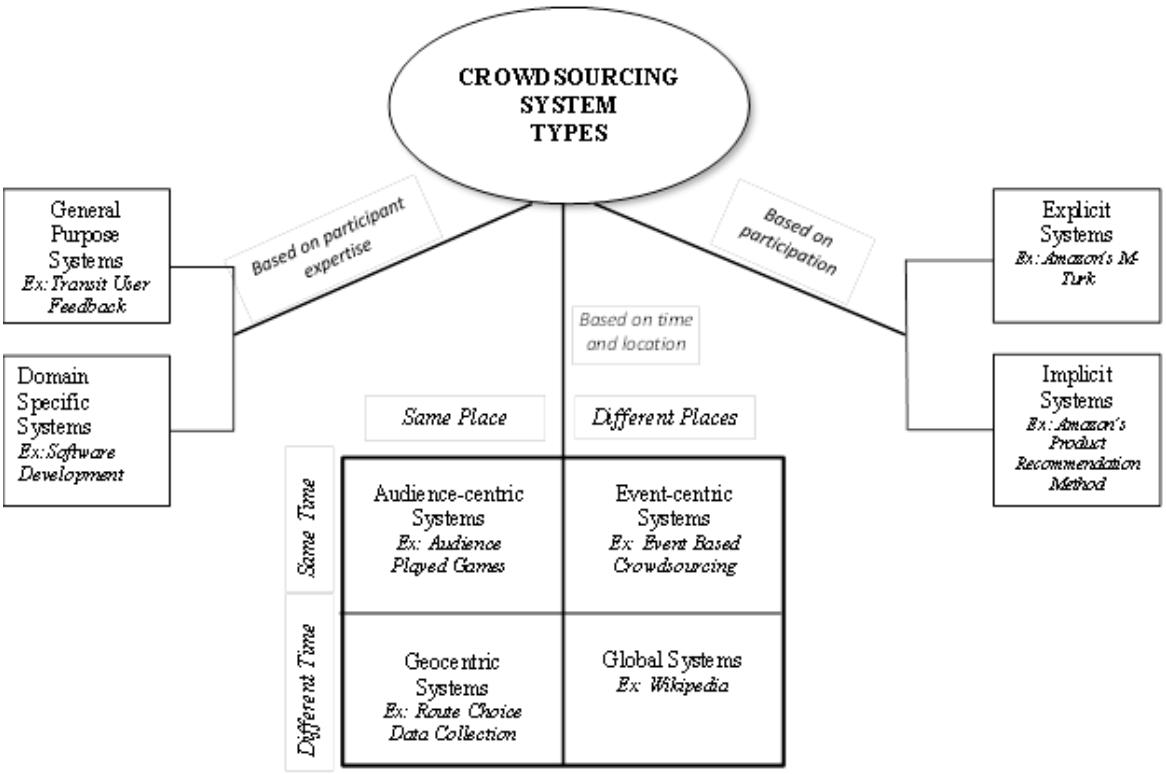


Figure 4. Classifications of Crowdsourcing Systems (Doan et al. 2011, Steinfield et al. 2013, Erickson 2010)

Steinfield et al. (2013) categorized public participation as either general purpose or domain specific systems. General purpose systems do not require any special expertise from the contributors and are not targeted to any user group in particular, while domain specific systems are designed for a special purpose user group (Figure 3). For example, most crowdsourced service quality feedback does not require any special expertise on the part of the participants and are hence, general purpose systems. Conversely, developing or beta-testing open source software through crowdsourcing requires expertise in particular programming languages and platforms and are hence, domain specific systems.

Crowdsourcing systems are further classified based on whether the system is local or global in scope and whether the system is time bound or not (Erickson 2010) (Figure 3). For crowdsourcing systems where the participants are at the same place at the same time, the system is termed as audience-centric (e.g., clickers used in class discussions). For systems where participants can be at different places but the crowdsourced event is time bound i.e., it has a start and end time between which the collaboration has to happen, such systems are termed as event-centric. An example of event-centric crowdsourcing is organized online brainstorming sessions triggered by an event and spanning over a limited period of time. Systems where collaboration can happen between people from different places and over an indefinite period of time are termed global crowdsourcing systems (e.g., Wikipedia). Finally, systems where people are at the same place but the crowdsourcing is an ongoing process are termed as geo-centric crowdsourcing – an example is bicycle route choice data collection for a city.

Crowdsourcing Issues

As crowdsourcing keeps evolving and gaining popularity, different and larger systems are being experimented with and the issues uniquely associated with the characteristics of the systems are gradually surfacing. For example, domain specific systems automatically reduce the crowd size by requiring some expertise from the participants while implicit systems have the issue of not having explicit participant consent in using their contribution for the actual purpose of the project. *A priori* understanding of the project characteristics and hence its category can often largely help in setting up plans early to overcome such issues. The last case study presented in this paper is one such example where instead of making the system domain specific, an expert group is used as data quality auditor. This helps in retaining a larger participant base as well as provides the necessary check on the usability of the data collected through a general

crowdsourcing system. As crowdsourcing gets applied to different domains, and as the scale and scope of crowdsourcing systems increases, additional techniques for addressing these system specific issues need to be developed based on the requirement of the projects.

Beside the unique issues of the systems, operation and maintenance of crowdsourcing systems in general suffer from four major issues – (i) how to recruit and retain the participant base, (ii) user capabilities, (iii) how to aggregate the information provided by the users and (iv) how to evaluate the contribution of the users (Doan et al. 2011). The problem of recruiting and retaining participants is a major issue in adopting crowdsourcing for any project. Depending on the purpose of the project, it is often important that feedback is obtained from users with particular skills or expertise. Furthermore, retaining participants is often important for understanding a trend over time – to allow the crowd's understanding of the problem to evolve throughout the process. The use of recurring campaigns and marketing strategies at frequent intervals (along with new releases of apps) is suggested where applicable so that people remain curious about the project and the developers can help maintain a participant base over time (Priedhorsky et al. 2007). Using incentives in the form of material benefits as well as acknowledgement of contribution in the form of gratification announcements at project sites make people feel encouraged to participate in the project and can help recognize diverse kinds of contributions from the crowd (Doan et al. 2011).

Dealing with user capability is an important issue in citizen science projects and in problem solving projects where participants are required to have some background to appreciate the assigned task. While participatory planning may not generally require special skill sets, in cases where the planning process targets a special group, it is important that the participants are aware of the specific problems of that group (e.g., planning for bicyclists' needs requires

presence of people who bike in that area so that the relevant problems and issues are brought up on the table). In such cases, the crowdsourcing process may be most successful if it is designed as a domain specific system – rather than a general purpose one – where specific tools and capabilities are made available to develop and maintain relevant user capabilities.

Problems with data quality and challenges with data aggregation are two important issues that often undermine the benefits of crowdsourcing systems. Regarding the importance of data quality, Heipke (2010) assessed that "quality issues have been a primary point of debate since crowdsourcing results started to appear". From that perspective, a degree of loose hierarchical authority is needed to ensure that the data is useful for its intended purpose. Additionally, aggregation of the data from crowdsourcing is often a complicated task given the volume of responses received from a diverse pool of crowd participants. Coping with data issues is either often labor intensive as large data sets need to be manually cleaned, or more cost intensive as complex data management systems and processes need to be put into place in an attempt to reduce sources of human error.

Evaluating the contribution of the user is commonly accomplished by setting up an automatic screening program to evaluate the validity of user-submitted information based on predefined criteria. The screening program rejects any input that does not follow the set criteria and thus only valid information is retained. However, this kind of automation is possible only in cases where the input is sufficiently normalized to be evaluated programmatically – in cases where the responses are descriptive or subjective, there needs to be a manual evaluation stage where each response is evaluated based on its potential contribution to the project. Such manual processes are labor and cost intensive and are prone to subjective biases of the evaluator, but are also much needed in order to ensure data quality for the project.

Crowdsourcing and Its Use In Transportation

The characteristic of crowdsourcing that makes it suitable and useful for transportation planning is that it voluntarily brings together a large group of people on the same platform to address common issues that affect them. The use of crowdsourcing works successfully for local purposes through localized knowledge and acquired experiences (Brabham 2009) because people in a region tend to identify themselves with the region where they live, work, and socialize, and are generally more interested in the systems that affect them (Erickson 2010).

A survey of existing transportation systems which use crowdsourcing reveals that the predominant purposes of using crowdsourcing in these projects are either data or feedback collection from the users. For example, one popular use of crowdsourcing is in collecting route choice data from bicyclists using the GPS functionality of the user's cell phone – such data are not readily available through the standard data collection procedures and designing a separate survey for a small population of users is often not cost effective for regional planning agencies. Crowdsourcing in this case helps the geographically dispersed and diverse population of cyclists to work together on a common interest without financially burdening the planning agencies. Similarly, crowdsourcing can also help in collecting feedback from a socio-demographically diverse range of users of any transit system that can be immensely useful for improving transit service quality and standards.

Transportation related crowdsourcing systems designed to date can be implicit or explicit standalone systems as defined by Doan et al. (2011) and discussed in the previous section. They may also be either geocentric systems where only local users are engaged or global systems where any person can contribute to the system. Extending the categorization of public participation as defined by Steinfield et al. (2013), transportation crowdsourcing systems may be

further classified as either general purpose or domain specific systems. General purpose crowdsourcing systems do not require any special expertise from the contributors and are not targeted to any user group in particular, while domain specific systems are designed for a special purpose user group.

Examples of transportation related crowdsourcing are presented below with reference to the above mentioned classification systems: the first group of examples focus on receiving feedback from users while the second group of examples focuses on use of crowdsourcing for data collection. A standalone example is provided at the end of the sub-section as it deserves special mention for its use of data quality editors to ensure data usability and validity and at the same time, maintaining a broad user base, thereby addressing one of primary challenges of crowdsourced data collection. The section is followed by a discussion on the advantages and disadvantages of crowdsourcing systems.

Crowdsourcing Case Studies

User Feedback Based Crowdsourcing Systems

Three seminal examples of general purpose user-feedback systems are SeeClickFix (http://seeclickfix.com), PublicStuff (http://www.publicstuff.com) and FixMyStreet (http://www.fixmystreet.com), all of which rely on public feedback about neighborhood issues and have been successful in mobilizing communities to take up the task voluntarily. While FixMyStreet is essentially for users to report road maintenance issues, the developers have a similar transit-based tool called FixMyTransport (http://www.fixmytransport.com). SeeClickFix and PublicStuff can be used to report "any non-emergency issue anywhere in the world that a user wants to be fixed" (seeclickfix.com), be it infrastructural or governance related. In SeeClickFix, users can also set up neighborhood watches where they monitor and report local

community issues which are then taken up by advocacy groups or elected officials, and solutions are proposed publicly. It is evident from the nature of the participation in these cases that no special expertise is expected from the users. It is interesting to note that the majority of the reported issues are local and community oriented in nature, reinforcing the concept that crowdsourcing can be successful in addressing local and regional issues, making it suitable for transportation planning.

Shareabouts is another example of a general purpose crowdsourcing system utilizing an innovative approach. Shareabouts *(*http://www.shareabouts.org*)* is a web-based system that uses maps to generate user feedback on preferred location of facilities and amenities. A few ongoing projects that use Shareabouts are (i) Chicago Bikeshare where people pin preferred bikeshare locations on the map provided, (ii) North Carolina Alternative Bike Route Plan where people can vote for preferred alternatives as well as mark any segment that they think might be an inappropriate alternative, and (iii) Philadelphia Bike Parking Survey where crowdsourced information is collected for estimating the bike parking capacity of the existing stations and plan for future expansion. In Boston, Street Bump (http://streetbump.org) is a mobile application that uses a smartphone's accelerometer to detect potholes and other street hazards as people drive around the city – the geo-located street quality data collected through crowdsourcing is automatically uploaded and integrated with the city's process for locating and fixing pavement quality issues.

A transit project using a general purpose crowdsourcing system, OneBusAway was created to address the reliability issues with on time performance of transit systems in Seattle and to expand upon existing transit tools in the region. OneBusAway provides several feedback mechanisms (email, Twitter, blog, bug tracker) that allow users to make comments or

suggestions about the tools (Ferris et al. 2010). The design of the various tools, along with development of new features, has been further shaped by feedback from users via several user studies and the IdeaScale feedback platform (another general use tool that can be applied to transportation). Because OneBusAway is open source software, users have also submitted improvements of their own to the code. Thus, users eventually become partners in development and design of the OneBusAway program, which promotes a sense of community among the transit riders in the region and a sense of ownership of the program. This ownership is an important factor in maintaining the user base for the program (Ferris et al. 2010).

Another general purpose crowdsourcing project related to transit systems is Tiramisu transit (Zimmerman et al. 2011), a user feedback based real time information system for public transportation in Pittsburgh. Tiramisu Transit, a 'crowd-powered transit information system', uses riders as the human equivalent of automated vehicle location (AVL) thereby providing an innovative alternative to more traditional cost intensive data collection. Tiramisu Transit is a smartphone app developed by researchers at Carnegie Mellon University to improve users' transit experiences and transit accessibility (Zimmerman et al. 2011) Upon activation, the app shows a list of buses or light rail vehicles scheduled for arriving at that time – this list is based off past arrival data as well as real time data sent by riders on the vehicle. Tiramisu provides an option for the rider to indicate the level of fullness of the bus, which aids people with disabilities to choose the bus they want to access. Once aboard, the rider can use Tiramisu to find out which stop is next and to report problems, positive experiences and suggestions. Use of Tiramisu is motivated by the rider's ability to use the same real-time arrival and fullness information they are reporting.

Crowdsourcing Systems for Data Collection

While issue reporting crowdsourcing systems like SeeClickFix and FixMyStreet do not call for any specific expertise from the user, there may often be systems where data and information are needed from a group with specific expertise or purpose, termed domain specific systems (Erickson 2010). Domain specific systems may be nested under a general purpose system, such as the bike projects undertaken using ShareAbouts. While all of these projects use the same crowdsourcing platform, the information is collected for one specific region, because it is more useful if it comes from the cyclists who use the facilities on a regular basis. Examples of standalone domain specific systems are the crowdsourced bike route data collection projects undertaken in San Francisco, Minneapolis, Atlanta, and Austin. These projects focus on developing smartphone apps and websites for cyclists to record their trips so that region-specific bikability maps can be created and facilities can be constructed on route segments as required.

CycleTracks (Hood et al. 2011) and Cycle Atlanta *(*www.cycleatlanta.org*)* are both projects for collecting bike route choice data through GPS enabled smart phones. The creation of CycleTracks by the San Francisco County Transportation Authority (SFCTA) in late 2009 was motivated by the lack of data on cyclists, cycling infrastructure, and eventually cyclist route choices. Traditionally, such data would be collected through public meetings because cyclists represent only 1-2 percent of commuters making vehicle count methods less useful. CycleTracks made participation in data collection for cyclists more accessible by moving data collection to the increasingly common smartphone use. In CycleTracks, first time users are asked optional information to determine cycling habits, such as riding frequency, age, gender, and zip codes for home, work, and school. Users record their trips by starting the app when they set out on a ride and then saving and uploading their data once they've reached their destination. The app records

bicycle trip route, time, distance, and average speed, along with user-reported trip purpose and notes. The trip data are wirelessly uploaded for analysis of cyclist route choice and is later used for planning facilities along the predicted routes (Hood et al. 2011).

Cycle Atlanta, a similar smartphone app for collecting data about cyclists and their routes within the city of Atlanta, was built off the open source codebase of the CycleTracks app. Cycle Atlanta also uses the GPS capabilities of smartphones to save and upload routes to provide basic data on how cyclists navigate the city, but the project team added features to the app including the ability to note with photos and textual descriptions of specific locations as either issues (pavement issues, traffic signal, enforcement, etc.) or amenities (bike parking, public restrooms, water fountains, etc.). The app also includes the collection of additional demographic data including cyclist ability and history as an indicator of comfort level to allow analysis of route data around an established taxonomy of urban cyclists (Dill and McNeil 2013), and to enable correlation with existing cyclist count and census data. As a distinctly different approach from CycleTracks, Cycle Atlanta categorizes cyclists into groups based on their cycling comfort level. The categories include strong and fearless, enthused and confident, comfortable but cautious, and interested but concerned. This categorization helps in understanding the preferences of different types of cyclists in choosing routes and hence can be immensely informative in creating a tailored application like bike maps for any particular group of users. Since the apps were launched in early October 2012, Cycle Atlanta has been used by over 1500 cyclists in Atlanta who have recorded more than 20,000 rides – represented by over 30 million individual data points. These data are the core piece of the City of Atlanta's effort to facilitate more streamlined communication between planners and cyclists.

A significant role of domain-specific crowdsourcing is in providing information from an otherwise unrepresented or underrepresented community. For example, due to the small size of the cycling community, bicycle maps are not commercially attractive and hence, are rare. Therefore, crowdsourced maps and geowikis are particularly suitable for understanding bicycle routes and for developing bicycle route maps (Masli 2011). Also, cyclists can benefit from regularly updated information, which is easy to maintain through "delegated responsibility among a motivated community with common purpose" (Masli 2011). Cyclopath *(*http://www.cyclopath.org*)*, a crowdsourced geowiki-based bicycle map developed by researchers at the University of Minnesota, provides an example of a domain specific use of crowdsourcing in transportation. Cyclopath maintains an active database of user-contributed bicycle routes and trails within the Minneapolis – St. Paul metropolitan area. The users of Cyclopath can add, modify, and delete roads and bike trails, segments thereof, points of interest, and neighborhoods. In addition, Cyclopath allows users to add notes and tags describing any feature on the map, such as 'bumpy' or 'closed'. Revisions are public and tagged to user logins for transparency and accountability. Cyclopath also has features that help the community to moderate itself. A list of 'Recent Changes' is also maintained, so that other users can identify and undo malicious modifications to the geowiki. Finally, Cyclopath allows a user to rate bike routes on a five-point qualitative scale (excellent, good, fair, poor, and impassable) for their own use and for aggregation to enhance bikability ratings. The Cyclopath community has made more than 13,000 revisions since release *(*cyclopath.org*)*.

Standalone Crowdsourced Data Quality Auditor System

Along with generating data from underrepresented groups, domain-specific crowdsourcing also helps in data quality management, which is an issue with self-reported data

in crowdsourced systems. As a study by Wiggins and Crowston (2013) revealed, most of the systems that use voluntary public participation include some form of expert control over the data. An expert user group can act as a bridge between general users and the system by filtering required information from general information and then by translating back the feedback from the system to the general users in a meaningful way. This helps in maintaining a feedback loop that is important in retaining participants and also prevents losing the critical mass which is often the case if the entire process is domain specific.

A standalone example of such an effort in transportation systems is the transit ambassador program initiated by the OneBusAway, Seattle program (Gooze 2013). The transit ambassadors are a *super user group*, with a solid understanding of the transit network and basic computational and analytical skills. Their role is to filter the incoming general purpose crowdsourced information and channel it to the respective departments within the transit agency for necessary action. Three core goals of the program development included addressing problem resolution, engaging the community, and improving agency-rider communication. Beginning in the fall of 2011, a number of errors with the real-time transit prediction data surfaced, affecting over 77% of a survey of riders (Gooze et al. 2013). While the OneBusAway mobile application included an error reporting function to allow users to identify errors experienced, the amount and quality of the crowdsourced reports began to overwhelm the OneBusAway administrators. Oftentimes, reports were duplicates of previously reported errors or the information submitted was incomplete and required additional effort to utilize it. With upwards of 500 errors reported on a weekly basis, the time required to evaluate these reports and any attempt to leverage them in order to resolve underlying problems with the real-time system would have required an effort from a collection of individuals. In contrast to previously described crowdsourcing programs,

this was not an issue of data collection, but rather a problem with information management. The management of the errors required the coordination between the agency, the OneBusAway administrator and the riding community; however, due to the constrained resources of each organization, there was no single contact to coordinate between these entities. This role fell to a collection of volunteer super users, or OneBusAway Transit Ambassadors. Figure 5 provides a visual summary of the flow of information established within the program and the role of the Ambassadors in coordination of the process.

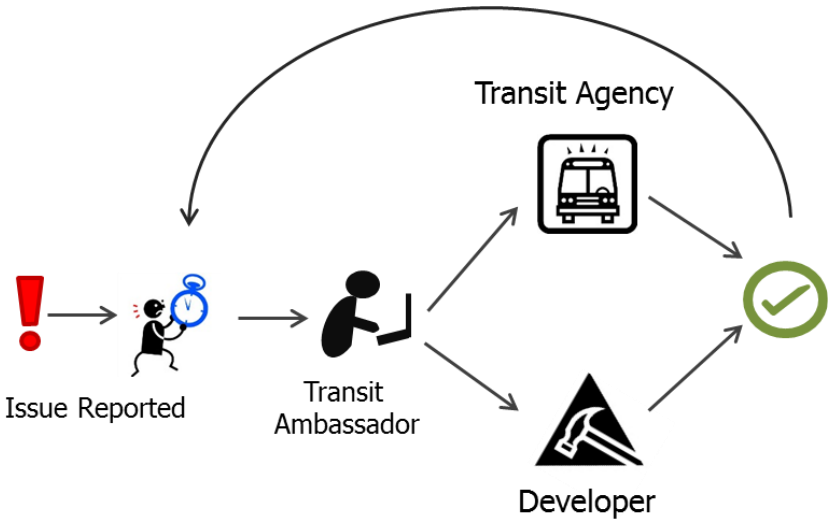


Figure 5. Information Flow of the Transit Ambassador Program

An initial group of three Transit Ambassadors were recruited from the rider community via blog solicitation and email outreach. The Ambassadors were provided resources such as transit schedule data, agency alert information and an "Error Decision Matrix" to assist in categorizing the crowdsourced error reports. All error reports were collected into an online database that allowed the Ambassadors to not just validate the error but to identify the nature and possible cause. This action of validation was a necessary and vital step in transforming the overwhelming amount of crowdsourced information from varying noise into usable knowledge.

Finally, the Ambassadors aggregated the information to forward onto the transit agency a clear and concise summary of notable issues reported by riders. For example, the summary of errors by vehicle and route provided the transit agency with valuable supporting information to help target potential actions to improve the real-time information system. The overarching role of the Ambassadors provided a level of expertise that could accurately evaluate the incoming error reports and thus efficiently triage and divert any relevant issues to the appropriate organization.

Providing a behind-the-scenes look at the underlying issues confronting the transit agency allowed the Ambassadors to relay that information to the rider community and to provide some context to the errors that everyone was experiencing. For example, a typical public relations response by the agency would have been interpreted far differently as compared to the Ambassadors relaying this information out in the community, which provided an enhanced level of trust. While some underlying real-time issues could not be resolved by the agency, the Ambassadors provided a means to explain to riders why an issue could not be fixed and how they could best adjust to the situation.

The success of the outreach exhibited by the Ambassadors and their role in representing not just the agency but the riders themselves gave validity to the potential that a fully deployed Ambassador program has within any real-time information system. With the proper adjustments to the available agency support and an expansion of the amount of Ambassadors, a Transit Ambassador program can effectively accomplish the core objectives and serve as not only a means for improving the real-time information product but serve as a mechanism for an agency to fully engage its riding community in a method that improves the overall functionality and quality of the transit service provided.

Summing It Up

Despite the fact that crowdsourcing has been used in transportation planning only recently, it is evident from the case studies presented that it has immense potential in augmenting/replacing traditional survey methods, particularly for groups of stakeholders who have a small user base in the transportation system. As seen with all systems, crowdsourcing also has its own issues that need to be addressed through proper planning and understanding of the system. Although there are criticisms with respect to data quality and data management issues, it is undeniable that crowdsourcing has been successful in engaging groups of people in solving a problem that affects their community. Crowdsourcing for bike route choice data has successfully solved the issue of data aggregation, defining a role of the users and linking their contribution to the final goal of the project by developing facilities for the bicyclists in San Francisco, Minneapolis, Atlanta and Austin. Meanwhile, transit information systems like Tiramisu Transit and OneBusAway have been very successful in redefining the role of their users in monitoring service standards and quality. The OneBusAway transit ambassador program has the potential to address the data quality issues associated with crowdsourcing by filtering and validating the data received from participants before the data reach the agency.

Most of the crowdsourcing systems use devices and technologies that are readily available and low cost – often crowdsourcing is based off devices that are owned by individuals (as in cycling data collection in CycleTracks and CycleAtlanta), involving no major financial investment on the part of the system. In an exemplary case, the Tiramisu project, described previously, uses crowdsourcing to actually replace the requirement of high cost AVLs. Tiramisu provides an example of ideal civic engagement in transit planning and operation where riders take care of other riders without the direct involvement of the transit agency and create an

information sharing legacy that is beneficial to both the users and the agency. With current funding limitations, crowdsourcing can be a preferred alternative to involve the public despite limited resources.

It should be noted however, that CycleTracks and CycleAtlanta are based on the widespread popularity and reach of the smartphone technology for crowdsourcing. While smartphones are easy to carry powerful devices that provide an inexpensive means of data collection, the usage of smartphones is restricted among some groups, such as those above 40 years old – thus, using smartphones for data collection comes with the issue of bias towards the input from these populations (Windmiller et al. 2014). Further research into possible biases arising from smartphone data collection is underway (Windmiller et al. 2014) and preliminary results show that age, income and ethnicity are the major factors that should be considered in smartphone data collection. This, however, can be addressed using proper outreach efforts and using supportive traditional methods for people who are not currently smartphone users.

Conclusion

Crowdsourced transportation projects bear evidence that crowdsourcing has the potential to bring together a large group of people on the same platform when there is an issue that affects them all. Systematic use of information and feedback from users for the purpose of transportation planning or for improving service standards is receiving significant attention recently and smart technology based crowdsourcing provides an ideal platform for engaging a broad group of users with limited additional financial burden on the system or the agency – possibly even replacing costly equipment. Crowdsourcing for data collection is found to be financially most effective in cases where the user base is small but enthused and motivated as in the case of bicyclists – in such cases crowdsourcing has a huge potential in augmenting the

standard data collection procedures by including the requirements of the otherwise marginalized groups of users. Examples of a few potential transportation related cases where crowdsourcing can be used are traffic data collection, getting user feedback for different systems, pavement and sidewalk quality monitoring and in understanding people's opinion in creating new facilities.

Crowdsourcing issues are mostly concentrated around problems with data quality, accuracy and data aggregation. However, these issues may be addressed through proper planning and with an understanding of the final goal of the crowdsourcing project. Further research and implementation of such strategies in real life projects are needed to establish a generic framework of crowdsourcing for transportation planning.

CHAPTER 3

SOCIO-DEMOGRAPHIC INFLUENCE ON RIDER TYPE CLASSIFICATION AND INFRASTRUCTURE PREFERENCE

Introduction

Popular adoption of bicycling as a mode of transportation can reduce overall congestion, air pollution and energy consumption while at the same time enabling an active lifestyle and providing users with a low cost, equitable means of transportation (Sallis et al. 2004, Damant-Siriois et al. 2014). In view of all these prospective benefits, the federal government has recently reoriented its policies towards promoting biking and walking (FHWA 2009). Additionally, several state and local transportation planning agencies have incorporated a bicycle planning module in their long term vision for their regions, including Atlanta (www.atlantaregional.com/plan2040). Despite this recent interest, research shows that although 40% of the trips made in the U.S. are less than 3 miles and may therefore be a bikeable trip due to the short distance, only 1.8% of such trips are bicycle trips (Pucher et al. 2011). This low usage of bicycling has been generally attributed to safety issues (AASHTO 2012), with major safety perception factors including high speed limits, high traffic volumes, last mile disconnect in the network, and an absence of physically separated facilities for cyclists (Dill and Carr 2003, Buehler and Pucher 2012).

Studies reveal that a substantial increase in the number of bicyclists can be achieved by providing facilities for safe riding (Pucher and Buehler 2009), and therefore it is important for planning agencies to know where cyclists prefer to bike and their desire for dedicated facilities. Cities often try to organize the route network by balancing the connectivity of the network, shortest travel distances, parking locations, and traffic volumes (Dill 2004) – but this task is

often difficult as perception of safety and comfort may vary across the level of experience of the cyclists, age, gender, traffic characteristics, and other factors. This difference in perception is further complicated by a lack of data on individuals who are not cycling at present, but who may otherwise choose bicycling as a mode if proper infrastructure and environment are provided. Geller (2006) suggested that the majority of Americans belong to this group, and that large scale adoption of bicycling for transportation is dependent on making bicycling a viable and acceptable option for such riders. He hypothesized, based on his experience working as a bicycle planner with the City of Portland, that infrastructure preferences are different across the population of Portland cyclists. According to Geller (2006), these preferences are reflected in their level of comfort and willingness to bicycle given certain combinations of road characteristics and bicycle infrastructure.

Multiple studies have attempted to group the diverse range of cyclists and their perceptions into categories based on person-level attributes as well as trip-level attributes so that preferences and perceptions of cyclists and non-cyclists can be predicted even when data is sparse. In addition, there have been studies relating socio-demographic characteristics to cycling propensity, although results do not always agree. Being female appears to be the only consistent attribute which shows a negative propensity towards cycling (Krizek et al. 2005, Gerard 2006, Gerard et al. 2007, Emond et al. 2009, Akar et al. 2013, Segdahlia and Sanchez 2014b). However, even that result holds true only for cyclists in the U.S. and Australia – the proportion of cyclists that are female is higher in the Netherlands and Denmark, while female cyclists in Germany make more trips on average than their male counterparts (Gerard 2006). Studies by Gerard et al. (2007) and Krizek et al. (2005) point towards a greater safety concern and absence of cycling infrastructure as reason for this lower adoption rate of cycling among females in the

U.S. and Australia. They suggest that aversion to cycling in absence of cycling infrastructure may stem from the more risk averse nature of women than men under similar situations. On the other hand, drawing parallels from health behavior research, Emond et al. (2009) suggested that there are multiple attitudinal factors that influence propensity to bicycle including self-efficacy or a person's ability to confidently engage in an activity.

In this research, we use data collected from cyclists in the Atlanta region to answer two questions: (i) whether self- efficacy, described here as the level of comfort and confidence in bicycling, varies across socio-demographic attributes of the cyclists and (ii) if infrastructure preference is influenced by the confidence and/or socio-demographic attributes of the cyclist. Accordingly, the analysis is presented in two parts: in the first part, we use different logistic regression models to understand the dependence of self-classified rider type on socio-demographic variables and riding characteristics like riding frequency and riding history. In the second part, we use multiple stated preference survey datasets to understand if there are any preferences for infrastructure type (bike lane vs bike path) or road characteristics (slopes or traffic conditions) that vary distinctly across the rider types. The survey data were analyzed using factor analysis and regression models were constructed to understand the correlation between confidence levels and infrastructure preferences modeled as different factors.

Literature Review

One of the earliest studies on route and infrastructure preference of cyclists was done by Aultman-Hall et al. (1997) for the city of Guelph, Ontario, Canada. The user data relied on both a mail out community survey and a user survey distributed to cyclists at cycle shops and on the road. The study revealed that men bicycle for statistically significantly longer distances than women. Cyclists were found not to choose an off road path or trail until the quality of such road

or trail was significantly better than the city streets. Collector roads were universally favored over arterials, although speed limit of the link did not seem to influence route choice. For the following characteristics, a statistically significant difference was found between the chosen route and the shortest route: (i) grade – cyclists prefer routes with less grade than a shortest route would have; (ii) traffic signals – cyclists tend to use routes with more signals than the shortest route and they particularly use traffic signals for any turning movement; (iii) number of turns – cyclists seem to prefer routes with fewer turns than the shortest route, and finally, (iv) cyclists tend to avoid links with more than two buses per hour.

Another stated preference survey study, conducted in St. Paul – Minneapolis by Krizek et al. (2006), revealed that cyclists prefer on-street bicycle lanes to off-street trails. Streets with bike lanes were also found to be preferable to streets with no on-street parking but no bicycle lane as well. Tilahun et al. (2006) conducted an adaptive stated preference survey to determine cyclists' preference between off-road facilities, designated bicycle lanes with parking and with no parking, and shared facilities with varying levels of traffic. The results indicate that higher income households have higher odds of choosing better facilities, and age and sex are not significant, although females are more likely to choose safe facilities than men.

Sener et al. (2008) used a web-based stated preference survey to collect data on cyclist route choice in Texas and used a panel mixed multinomial logit model to estimate route choice parameters. Of the chosen attributes, they found that cyclists expressed a sensitivity to travel time and preferred streets without on-street parking. Moderate hills were preferred over flat terrain and cyclist characteristics entered the model specification only as interaction terms between parking and cyclist age – no significant influence of any other cyclist characteristic (experience, for example) was noted. Titze et al. (2008) performed a study on 1000 bicyclists in

Graz, Austria where the participants were asked to fill out a survey related to physical environment, home and destination environment, social environment, living quarters and attitudes. Their study reported that presence of bike lane connectivity, social support, and perceived benefit of rapidity have a positive influence on bicycling while perceived barriers of impractical mode of transport and physical discomfort act as major deterrents to bicycling. Xing et al. (2008) conducted an online survey among residents of five cities similar to Davis, California. Using individual level factors such as socio-demographic and attitudinal attributes, social environment factors such as the perception about bicyclists and bicycling, and physical environment variables such as presence of bicycle infrastructure, they sought to understand how these factors influenced owning and using a bicycle. The study showed that attitudes such as "I like biking" are significant in influencing a decision to own or use a bicycle. Similarly, a social perception of who the cyclists really are can significantly impact the possibility of owning and using a bicycle. Emond et al. (2009) studied the influence of gender on a binary dependent variable indicating whether the participant bicycled in the last 7 days or not. They used the same ecological model and the same variables as those used by Xing et al. (2008). The results indicate that there are significant differences in attitudes and preferences between men and women cyclists. While male cyclists find living in a bicycle-friendly community to be a positive reinforcement and are more likely to bicycle if they bicycled in their youth, female riders are mainly motivated by safety factors and are much more affected if the size of the street increases or if there are no bike lanes. Comfort is the most important factor for women bicyclists and their household responsibilities possibly deter them from bicycling, as reflected in the significance of the factor 'I need a car to do many of the things that I like to do' for women cyclists. Social perception (bicyclists are rich/ bicyclists are poor) and access to safe destinations influence both

genders similarly and so do socio-demographic variables like education and child assistance, both of which positively influence propensity to bicycle. However, as with Xing et al. (2008), 'I like biking' is the most significant positive factor for both groups of cyclists. The study by Emond et al. (2009) points to the fact that instead of experience only, gender should be an important consideration when planning for bicycle infrastructure. The present bicycle compatibility index (BCI) used by FHWA for planning purposes assumes that cyclists become confident with experience and will be comfortable bicycling under most traffic situations. Therefore, separate infrastructures are planned for inexperienced riders and mostly for recreational purposes which often fails to provide access to services. As the study by Emond et al. (2009) shows, female riders have marked preference for separate infrastructure irrespective of experience and hence, such facilities should be designed more along the primary road network. The study also attributed the gender difference in bicycling to difference in perceived safety of bicycling and the difference in comfort level across different facility types between men and women cyclists.

Winter et al. (2011) conducted a survey among 1402 current and potential bicyclists in Vancouver, Canada to understand the 'potential motivators and deterrents of cycling'. Respondents were grouped into potential (n =197), occasional (n = 617), frequent (n = 481), and regular (n = 107) cyclists. 73 potential motivators and deterrents to cycling were identified from literature and presented in the survey questionnaire as "how would [item X] influence your decision to cycle?" The responses could be marked on a 5 category behavioral intent scale with much less likely to cycle having an influence score of – 1 and much more likely to cycle having an influence score of +1 and intermediate categories being marked at increments of 0.5. From the mean response score, the top 3 motivating items were "the route is away from traffic noise & air

pollution", "the route has beautiful scenery" and "the route has bicycle paths separated from traffic for the entire distance" while the top 3 deterring items were "I need to carry bulky or heavy items", "the route has surfaces that can be slick when wet or icy when cold" and "the route is not well lit after dark". Safety factors, which included items like 'risk from motorists who don't know how to drive safely near bicycles' and 'risk of injury from car-bike collisions', had the highest factor scores as a deterrent to cycling has the highest negative mean factor score on influence on likelihood of cycling, followed by poor weather and darkness, interactions with motor vehicles and route surfaces. Mean scores and ranks were similar across the different groups. The factors conducive to cycling were ease of cycling, integration with transit, bike parking, and end-of-trip facilities. Lane marking and signage also scored substantially high as factors conducive to cycling. The same participants were provided with pictures of 16 different routes and asked to indicate which routes they would choose to cycle (Winter and Teschke 2010). Route preferences were found to be similar across all the cyclist types (potential, occasional, frequent and regular). Between 70% and 85% of the participants chose off-street paths, 70% of the participants would bike on physically separated routes next to major streets and 50-65% would bike on residential routes. Routes with bike lanes, paved surfaces, no on-street parking and traffic calming increased the likelihood of choosing the route from 12% to 37%.

Yang and Mesbah (2013) conducted a survey on a small sample at University of Queensland, Brisbane with 19 respondents, majority of whom were high school students and postgraduates. They found that distance, travel time, traffic safety and gradient are the most important factors in route choice of bicyclists. Akar et al. (2013) sent out an online survey for the student, staff and faculty of Ohio State University (OSU) through the OSU Transportation and

Parking Service's (T&P) webpage. From about 2000 respondents who provided nearly complete data, results indicate that female respondents are more likely to overestimate their commute distances, do not feel safe in vehicular traffic and are more deterred from cycling by the absence of bike lanes/paths/trails. Female participants also cited the need to change clothes and carry things as major reasons for not bicycling. A significantly lower percentage of female cyclists considered themselves advanced cyclists (9%) as compared to male cyclists (35%). Female cyclists are also significantly more likely to feel biking and walking on campus after dark unsafe – about 30% of female participants agreed that they feel safe biking and walking on campus after dark as compared to about 70% of male participants. Mode choice models indicate that being a female makes it less likely to be a cyclist, while having bike lanes and trails and feeling safe positively impact the possibility of choosing cycling for the commute.

Segadilha and Sanchez (2014a) conducted a survey on 49 frequent cyclists in the city of Sao Carlos, Brazil to identify the relative importance of multiple factors identified from the literature. Slope appeared to have the least importance in choice of route, while number of trucks, number of buses, traffic volume, and traffic speed score the highest in influencing route choice decisions. The authors also reported that stratification by age, gender, and cycling frequency showed that preferences may vary significantly across these categories. Sousa et al. (2014) conducted an email survey involving 380 students from 3 different cities in Brazil. The survey provided the participants with six statements on perceived barriers to cycling and asked them to rate how positively or negatively these barriers influence their decision to bicycle. The perceived barriers included absence of adequate infrastructure, traffic safety, distance, physical ability and experience, slopes, and climate. Lack of adequate infrastructure, lack of safety, and slopes were found to be the most important barriers to cycling, while climate was the least

important. Wang et al. (2014) conducted a survey at the University of Auckland to understand the factors that influence a cyclist's decision to bicycle and his/her route choice. The study concluded that safety, low traffic volume and speed, separation from cars and pedestrians, and separate facilities are important factors in promoting cycling, along with connectivity and ability to carry the bicycles on public transport. Wang et al. (2014) also concluded that safety is a more important factor for female than for male cyclists.

Additionally, Barros et al. (2015) designed an online survey to understand the different factors that affected mode and route choice. Their findings suggest that presence of cycle lanes and bicycle parking encourage people to choose bicycling. Mertens et al. (2015) conducted a web-based survey with 389 respondents where participants were presented with photographs of two alternative cycling routes. The study was designed to understand how macro-environmental factors like residential density interacted with micro-environmental factors like speed limit and presence of bicycle facilities to affect participants' decision to bicycle. Mertens et al. (2015) found that while participants preferred low residential density over medium or high residential density, preference for a low speed limit and physically separated bicycle facility does not vary across choice of residential density.

On another subject, several studies have developed classification systems for cyclists based on person-level attributes as well as trip-level attributes. Damant-Siriois et al. (2014) and Dill and Voros (2008) provide comprehensive accounts of different cyclist type classifications based on both person- and trip-level attributes. Person-level attributes used in classifications include the attitude and comfort level of cyclists (Geller 2006), the behavioral perspective and value system of the individual (Paulssen et al. 2011) and a cyclist's preference for infrastructure (Larsen and Geneidy 2011). Trip-level attributes include trip purpose (Kroesen and Handy 2013)

and whether trips depended on weather conditions (Bergstrom and Magnusson 2003). However, Geller's classification of cyclists into four different categories of *strong and fearless*, *enthused and confident*, *interested but concerned*, and *no way no how*, based on their comfort level in cycling, gained notable popularity and was used in planning for cycling infrastructure by multiple regional planning agencies in the last decade (Dill and McNeil 2012; Geller 2006). It should be noted though that this classification was devised on an ad-hoc basis and was not based on any survey or self-description of cyclists (Geller 2006). Subsequently, in a recent study, Dill and McNeil (2012) conducted a random phone survey of 908 adults in Portland, OR asking respondents about their comfort level in bicycling on non-residential streets, with and without bike lanes, to set up a basis for the categorization. This comfort level question was combined with an interest question which asked if respondents wanted to bicycle more than they are currently doing. The answers to these two questions were considered together to categorize riders into the classifications suggested by Geller (2006). The study categorized the riders who were comfortable bicycling on non-residential streets even without bike lanes as *strong and fearless* irrespective of their response to the interest question. Cyclists who were comfortable on non-residential streets only with bike lanes were classified as *enthused and confident* riders. The cyclists who were not comfortable on any facilities and/or have not bicycled for transportation for the last 30 days were categorized as *no way no how* while the *interested but concerned* group had cyclists uncomfortable on residential streets irrespective of their interest in bicycling.

New Cyclist Categories

According to Geller (2006), the *strong and fearless* riders are cyclists who would bicycle irrespective of road and traffic conditions and whether separate facilities are present or not; the *enthused and confident* riders are cyclists who will choose cycling as a mode of transport even if

bare-bones facilities are present or if it is not infeasible for them because of distance or road features; *interested but concerned* is the group of individuals who are currently very infrequent bicyclists or do not bicycle at all and will bicycle only when they have protected and separate facilities for the purpose i.e., these are the riders who are willing to bicycle if proper infrastructure is provided. The fourth category of riders, *no way no how*, includes the individuals who will not bicycle under any circumstances.

There are multiple ways in which this classification system can be improved. First, within a wide spectrum of cyclist 'types', the classification misses those people who are enthusiastic bicyclists but are not willing to bicycle with bare-bones cycling infrastructure. Most often, their concerns are more related to safety than confidence (for example, riding together with children). These cyclists possibly bike commute every day but using different routes than the enthused and confident group, and often undertake longer detours to find safer routes. On the other hand, while they prefer separate cycling infrastructure, these are the people who can also be motivated by traffic calming measures and do not require a physical separation from the traffic to be able to bicycle. Therefore, this group of cyclists belongs neither to the *enthused and confident* group, nor to the *interested but concerned* group as proposed by Geller.

Geller's classification also misses the captive riders of the system – the people who cycle because of a lack of alternatives. While these cyclists may make frequent and regular trips, they are not bicycle enthusiasts and generally associate a negative social image with cycling. It is much more difficult to retain such users in a cycling system without changes in the social perception that cyclists are either poor or rich and that cycling is not the natural normal mode of transportation (Xing et al. 2007, Emond et al. 2009). Efforts in that direction will involve not

only building infrastructure, but also creating awareness and education, thus requiring a different policy approach than that for other groups.

Our research primarily focused on collecting data from cyclists via a GPS enabled smartphone application. We assumed that non-cyclists will not use the application and hence, the *no way no how* category was not applicable. Therefore, for our research, we modified the existing rider type classification and added a new group called *comfortable but cautious* to the existing system. This new category was designed to adequately represent the riders who differ in their view of safety from more aggressive riders but, at the same time, are similarly enthusiastic about bicycling. We also assumed that most captive riders will not be motivated enough to provide personal and cycling data voluntarily and therefore did not create a separate group for them. However, future work should consider their preferences as well. In its final form, the rider types suggested in this research consisted of four different groups:

(i) strong and fearless;

(ii) enthused and confident;

(iii*) comfortable, but cautious;*

(iv) interested but concerned;

While the other groups are expected to show similar attitudinal preference as the Portland study (Geller 2006, Dill 2012), the *comfortable and cautious* group of riders is hypothesized to include a greater proportion of female cyclists and/or individuals in higher age groups who are bicycle enthusiasts, but are less risk-taking in attitude and hence may appear to be less confident.

Methodology

The primary purpose of this study is to develop a model that can help us to relate readily available socio-demographic data to cyclists' stated preferences for infrastructure. Toward that

end, we used multiple data sources and models to find the combinations of attributes that can best predict the infrastructure preferences of cyclists. First, we used socio-demographic data collected from riders who recorded their trips on a smartphone app called Cycle Atlanta. We hypothesized that the rider type classification can serve as a proxy for how cyclists of different ages, genders, incomes, and ethnicities perceive risk and comfort on the streets of Atlanta. We then used stated preference survey data from two online surveys administered separately by two different groups and at a time gap of six months. The first survey was conducted by the Atlanta Regional Commission across the region, and the second was conducted by our research group and geared to the users of the Cycle Atlanta smartphone application. The survey questions related to socio-demographic information and infrastructure preferences of the participants were carefully designed to ensure that they had identical wording and choice order in both the surveys. The data on infrastructure preferences were then analyzed using factor analysis to group similar, correlated preferences under one factor. The factor scores were then regressed against socio-demographic variables to understand how they influence a participant's infrastructure preferences.

Part 1. Predicting Rider Type based on Socio-demographics and Riding Behavior

In the first part of the analysis, we directed our efforts towards identifying the relationship between stated rider type and other socio-demographic variables of participants.

Data Source: Cycle Atlanta

The first analysis uses the data collected through the Cycle Atlanta smartphone application, developed through a collaboration between the Georgia Institute of Technology and the City of Atlanta's planning office to promote cycling in Atlanta (The City of Atlanta, 2011).

The application was named Cycle Atlanta after the larger planning project for which the application was initiated, and was developed by an interdisciplinary team of researchers. The application was originally based on San Francisco's CycleTracks (Hood et al. 2011), althoughCycle Atlanta was substantially updated to make better use of current features available in iOS and Android as well as to include features that the City and local bicycle advocacy groups wanted in the application. The basic feature is trip recording, where the application uses the GPS of the phone to record the location of the user once per second. In addition to tracking cyclists' trips, the app also provides options to enter personal information, including age, email address, gender, ethnicity, home income, zip codes (home, work, and school), cycle frequency, rider type, and rider history (Misra et al. 2014).

The breakdown of age, gender, income, and ethnicity was kept similar to the breakdown as found in the household travel survey. The age and income intervals as well as the gender and ethnicity subcategories were adopted from the household travel survey conducted by Atlanta Regional Commission (www.atlantaregional.com/transportation/travel-demand-model/household-travel-survey). The rider type and rider history categories are exclusive and unique to the design of Cycle Atlanta. The cycling experience field allowed users to specify how long they have been cycling and can choose from the categories 'since childhood', 'several years', 'one year or less' and 'just trying it/just started'.

As of June 2014, the Cycle Atlanta dataset consisted of 1529 unique users who could provide information on their age, gender, ethnicity, income, rider history and cycling frequency. Because there were only 6 cases in the age group of 65+, that group was merged with the age group of 55-64 years old and the new group is referred to as "age 55+" for the rest of the analysis. About 60% of the riders provided information on each of the socio-demographic

categories. The users of Cycle Atlanta are predominantly male (about 75%), white (about 80%) and mostly from a high income group (>$75,000) (about 45%). Table 2 presents the basic statistics of the different socio-demographic variables considered in this study. The median age of the users is between 25-34 years, while the median income is between $60,000 and $74,999. The median rider type is an *enthused and confident* rider with median cycling frequency of several times per week and a median riding history of several years.

Analysis and Results

The goal of this part of the study was to understand the relationship between cyclist self-classification into different rider types and the socio-demographic make-up and riding pattern of the cyclists.

Multivariate Analysis

Except for ethnicity and gender, the socio-demographic variables considered in this study have an underlying order, although they are categorical. This led us to use methods and analyses relevant to ordinal variables instead of nominal variables. To understand degree of association between variables, polychoric correlation was used, which assumes an underlying continuous bi-variate normal distribution for discrete categorical variables with an ordinal scale. Figure 6 shows the correlation coefficients obtained from the analysis. Age and income are correlated with a measure of correlation in the range of 0.5. Rider type is correlated with gender (being male), cycling frequency, and rider history, each with a correlation ~ 0.35.

Figure 7 shows the percentage of rider types across the different socio-economic variables as well as rider history and frequency. As hypothesized, a high percentage of people in the *strong and fearless* group as well as in the *enthused and confident* group are in the age group of 25-34 and 35-44 and are male. These two groups also have a greater representation from

high income groups, indicating that people in those income groups are possibly more confident and aggressive than those in other income groups. Cyclists with a history of less than a year are more represented in the *comfortable and cautious* and the *interested but concerned* groups than any other group while a high proportion of *comfortable but cautious* or *interested but concerned* riders are infrequent cyclists.

Table 2. Basic Statistics for Socio-demographic Variables

Socio-demographic Variables(n = 1529)		
Age(n = 1001)	Count	Percentage
Less than 18	6	0.6
18-24	110	10.99
25-34	448	44.76
35-44	218	21.78
45-54	144	14.39
55-64	66	6.59
65+	9	0.9
Gender(n = 981)		
Female	240	24.46
Male	741	75.54
Income(n = 776)		
Less than $20,000	78	10.05
$20,000 to $39,999	133	17.14
$40,000 to $59,999	111	14.3
$60,000 to $74,999	95	12.24
$75,000 to $99,999	112	14.43
$100,000 or greater	247	31.83
Ethnicity(n = 955)		
African American	46	4.82
Asian	43	4.5
Hispanic / Mexican / Latino	53	5.55
Multi-racial	22	2.3
Native American	3	0.31
Pacific Islander	2	0.21
White	770	80.63
Other	16	1.68
Cycling Frequency(n = 546)		
Daily	158	28.94
Several times per week	260	47.62
Several times per month	113	20.7
Less than once a month	15	2.75
Rider Type(n = 989)		
Strong & fearless	187	18.91
Enthused & confident	443	44.79
Comfortable, but cautious	333	33.67
Interested, but concerned	26	2.63
Rider History(n = 985)		
Just trying it out / just started	59	5.99
One year or less	120	12.18
Several years	330	33.5
Since childhood	476	48.32

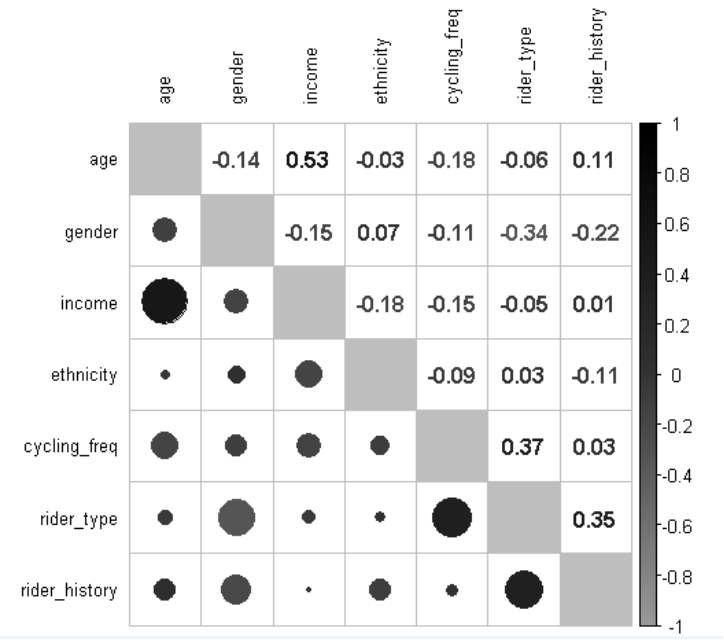


Figure 6. Measure of Association between Variables

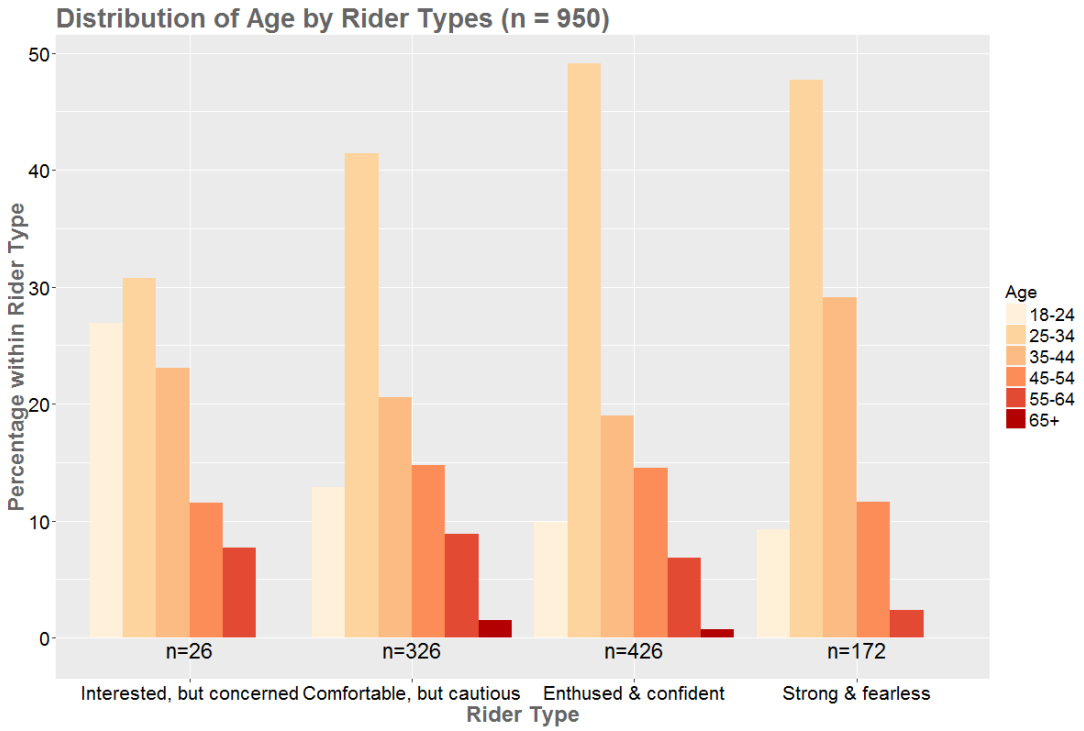


Figure 7(a). Distribution of Age by Rider Type (n = 950)

Figure 7(a) – 7(f). Socio-demographic and Riding Pattern Distribution of Cyclists across Rider Types

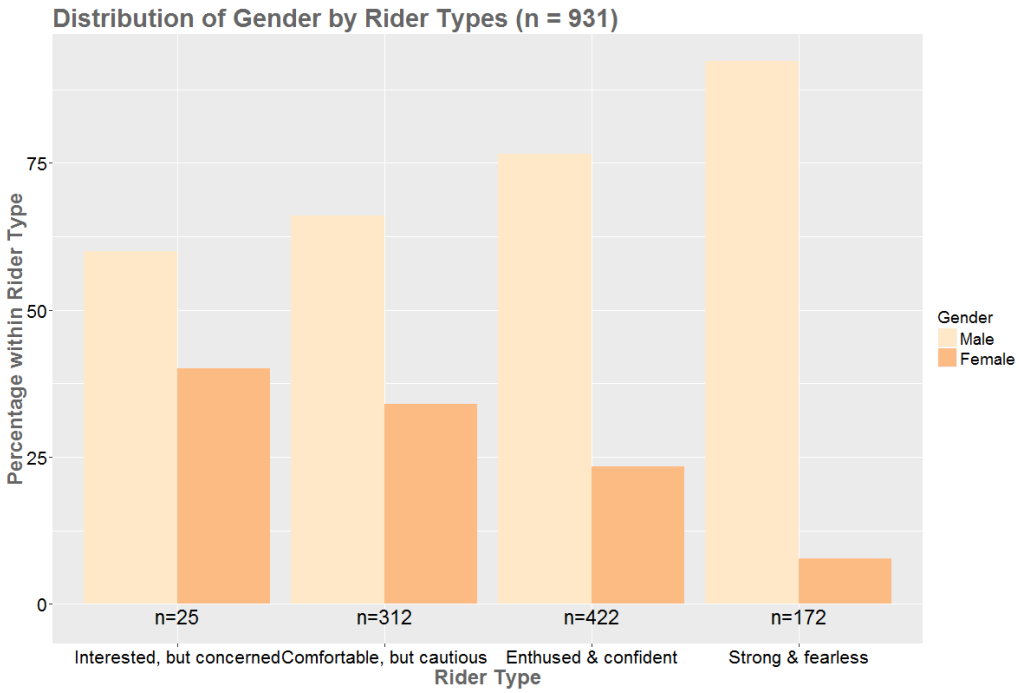


Figure 7(b). Distribution of Gender by Rider Type (n = 931)

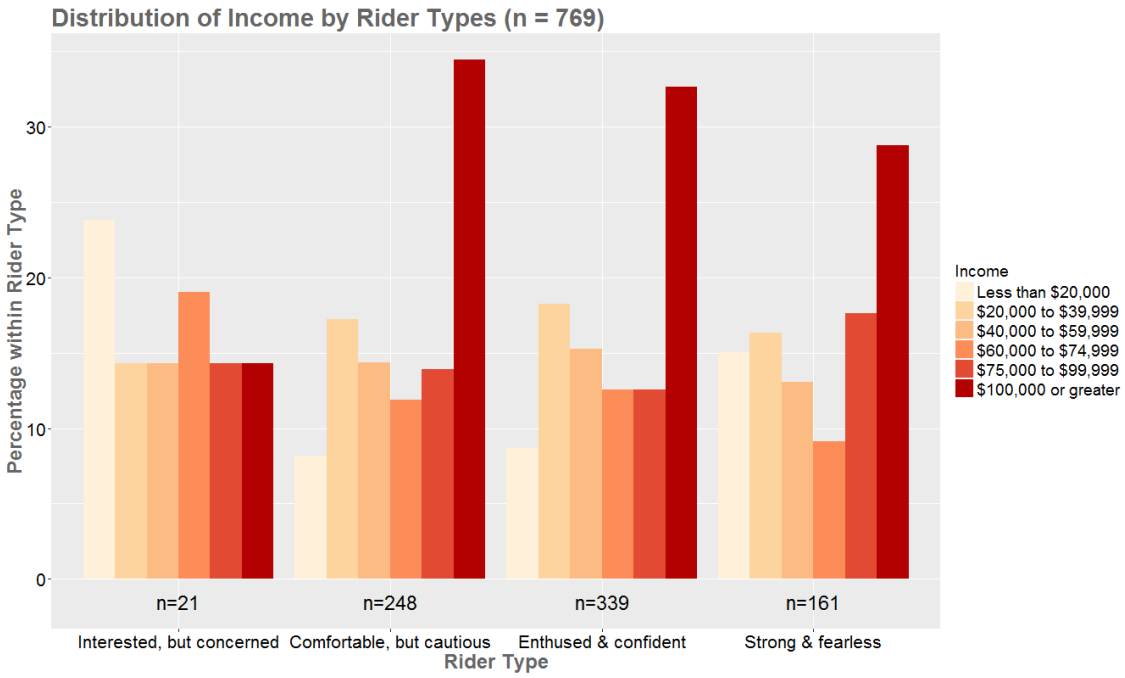


Figure 7(c). Distribution of Income by Rider Type (n = 769)

Figure 7(a) – 7(f) Socio-demographic and Riding Pattern Distribution of Cyclists across Rider Types Continued

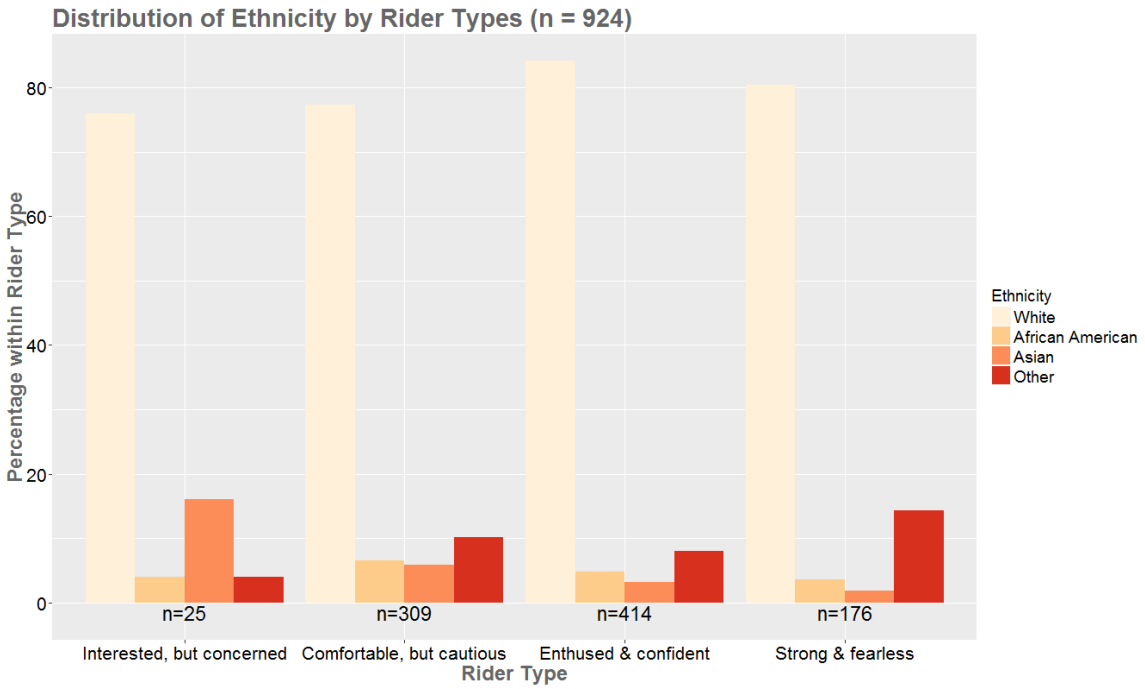


Figure 7(d). Distribution of Ethnicity by Rider Type (n = 924)

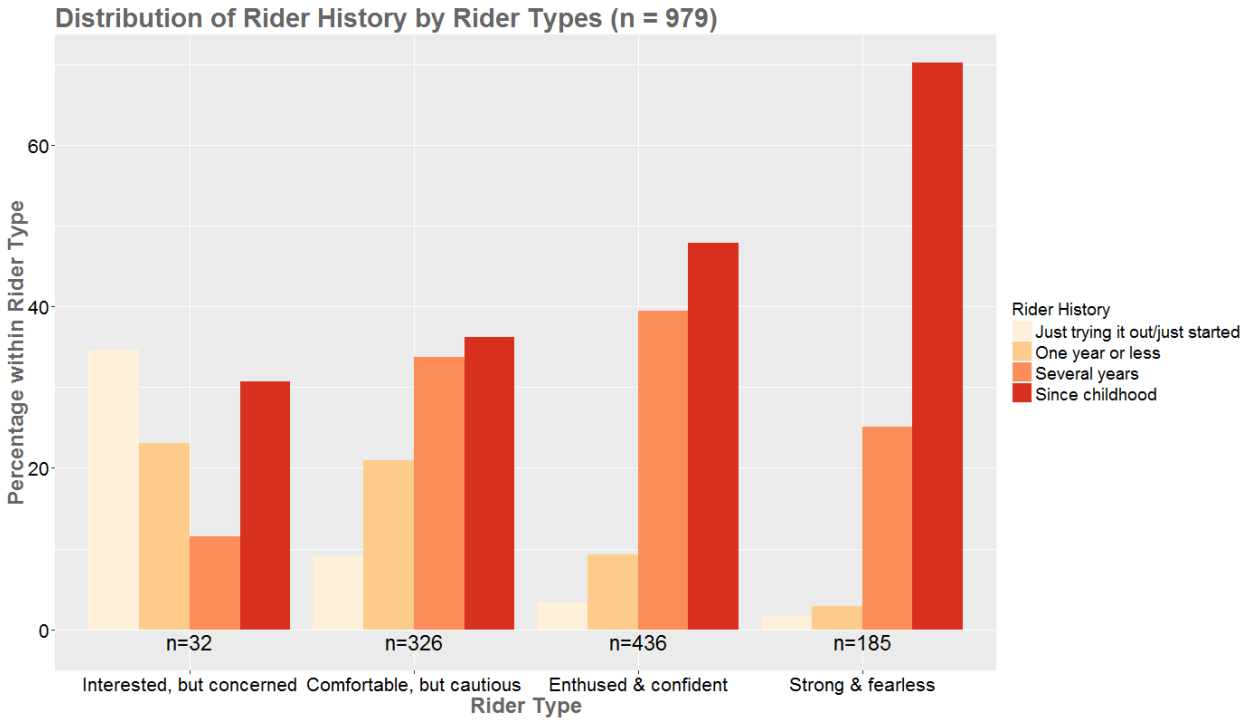


Figure 7(e). Distribution of Rider History by Rider Type (n = 979)

Figure 7(a) – 7(f). Socio-demographic and Riding Pattern Distribution of Cyclists across Rider Types Continued

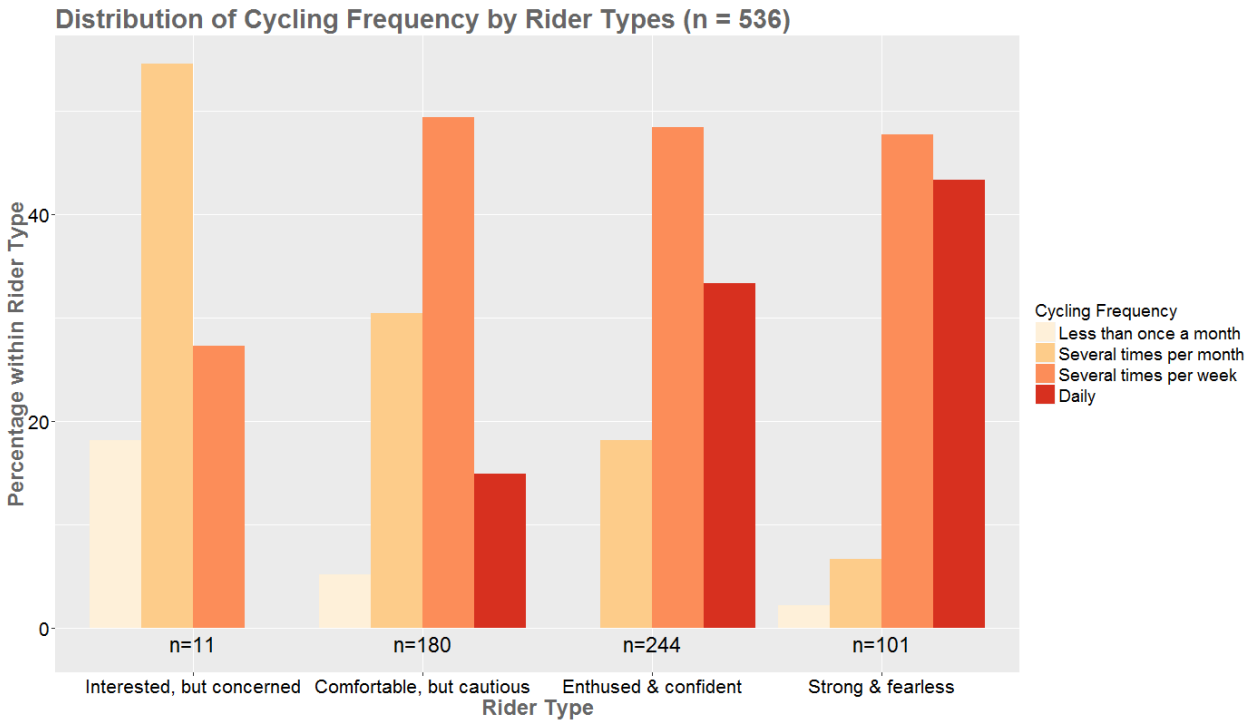


Figure 7(f). Distribution of Cycling Frequency by Rider Type (n = 536)

Figure 7(a) – 7(f). Socio-demographic and Riding Pattern Distribution of Cyclists across Rider Types Continued

Logistic Regression Models of Rider Type

Two main types of explanatory variables were used in these models – the socio-demographics and the riding habit/pattern of the participant. The socio-demographic variables included age, gender, income and ethnicity while the riding pattern variables included cycling frequency and rider history. From the distribution of age and gender across rider type, it was evident that there were very few participants in the age group above 45. So the age groups 45-54, 55-64 and 65+ were grouped into one category of 45+. The riding characteristics like cycling frequency and rider type pattern riding pattern was found to be distinctly similar across the age group of 25-34 years and 35-44 years and hence, these two groups were also merged to form a new group of 25-44 years. Similarly, different income categories were consolidated into 3

categories and different ethnicity types were consolidated into 4 categories. For rider history, the 'just started' category was merged with the 'less than once a week' category, resulting in 3 categories instead of 4.

Of the total 989 users who provided data on rider type, only 26 users classified themselves as *interested but concerned*. Cross tabulation of rider type across other variables showed *interested but concerned* riders having zero cell values with cycling frequency 'less than once a month' and small valued cells for age group 45+ (2 users) and ethnicity 'African American'(1 user) and 'Other'(1 user) thereby presenting a problem of quasi separation. Within cycling frequency also, there are only 13 users who have cycling frequency less than once per month and none of them are *enthused and confident* riders (0 users) which again presented the issue of separation. Quasi/complete separation implies a perfect prediction scenario where the dependent variable Y can be completely predicted by variable X when the separation is complete. In case of quasi complete separation, perfect prediction happens only for a subset of observations (Albert and Anderson 1984). For example, in this dataset, it can be predicted with absolute certainty that none of the riders who bicycle less than once per month will classify themselves as *enthused and confident*, although the same cannot be said about whether riders with cycling frequency less than once a month will classify themselves as *strong and fearless* or *comfortable but cautious*. Models estimated under quasi/complete separation are more likely to either not converge or give high co-efficient estimates and infinite standard error as the log-likelihood will be presumably flat (Zorn 2005). The most common way of dealing with quasi separation is to remove the problematic covariate which again might give specification bias if the covariate is strongly correlated. We ran models both by removing observations and by aggregating the sparsely populated group with its nearest neighbor. In case of cycling frequency,

the last group, cycling frequency less than once per month was merged with the group which bicycles a few times per month and the new group was named cycling frequency once or less per week. Models run by removing the observations with cycling frequency less than once per month gave a much lower model fit than the aggregated models and hence, in this paper, models with aggregated data are presented. Similarly, for addressing the quasi separation problem related to rider type, two alternative model sets were designed – one where the *interested but concerned* group (26 users) was merged with its next higher group *comfortable but cautious* (333 users) and another where the *interested but concerned* users were removed from the sample and models were estimated for the remaining three categories. The model estimates in either case were not significantly different and in keeping with our aggregation theme, in this thesis, the aggregated models are presented.

Based on these rider type distributions, logistic regression models were estimated for each rider type to understand how the self-described confidence level is affected by socio-economic variables as well as riding patterns of the cyclists. Several logistic regression models were explored to find the best way to represent the pertinent relationships. Since cycling frequency and rider type may have bi-directional causality, they were tested for explanatory power and likely association. A single variable ordinal model for rider type with cycling frequency as the explanatory variable gives a McFadden's ρ^2 of 0.48 but an ordinal model for cycling frequency with rider type as the explanatory variable gives a McFadden's ρ^2 of 0.07 (both unadjusted for sample size difference). Although it was found that cycling frequency has a greater explanatory power for rider type than rider type has for cycling frequency, in view of the simultaneity issue, models with cycling frequency and models without cycling frequency are both presented here.

Since the discrete observed rider type categories (y) were originally thought of as representing a latent continuous scale of confidence and comfort (y^*), two variations of the user's underlying decision process along that one dimensional scale were initially estimated. The first is where the self-classification process was thought of as representing a binary choice for each rider type (for example "Am I *strong or fearless* or not?"). This process was estimated using binary logistic regression models where the rider classifies himself/herself into a category ($y = 1$) if he/she perceives himself/herself above a certain confidence level threshold $(y^* > \tau)$; if the perceived confidence level is at or below the threshold $(y^* \leq \tau)$, the rider does not choose that rider type category ($y = 0$). Four different binary logistic models were estimated – one for each rider type. For each of these four choices, several models were run with different variable combinations to balance model fit and parsimony. Age group 45+, gender male, income less than $40,000, rider history since childhood, and cycling frequency of daily were chosen to be the base categories for age, gender, income, rider history and cycling frequency variables. Ethnicity was not included in the models due to its heavy bias towards white riders. Model fit statistics were calculated based off the corresponding equally likely model statistics (Mokhtarian 2016). In addition, even when not significant, variables with *t-statistic* >1 were kept in the models.

The first models were run with age and gender as explanatory variables which gave model fits in the range of 0.2 - 0.3 (with base equally likely). Age group 25-44 and gender were significant for *strong and fearless group* and for the group including *comfortable but cautious and interested but concerned*. At the second stage, income was added to age and gender. While income itself was not significant, McFadden's ρ^2 for these models ranged between 0.3 – 0.45 although the sample size reduced to 932 from 742. Walden's t-test did not show significance of the variable income (p = 0.94, 0.32). Since the correlation between age and income was earlier

found to be high (0.53), at the next step, an interaction term between age and income was introduced in the model. However, the model fit was not found to be significantly different from the previous model. In addition, introduction of interaction term led to perverse signs for the income variable. Therefore, age and income were included in the model as separate variables. Since models with age and income gave a better fit, we tested these models for multi-collinearity effect. The VIF (Variation Inflation Factor) test was performed on a linear version of the models and the VIF was found to be less than 5 for all variables including income.

Rider history was added to the model at the next step and was found to be significant across all the models. Wald's test as well shows that rider history is a significant variable ($p = 3.2$ e-09) for the model. At this stage, the ρ^2 values for the models range between 0.4 and 0.5 and both age groups and gender are significant across the *strong and fearless* and the *comfortable but cautious* and *interested but concerned* group. Rider history is the only significant variable for the *enthused and confident* group at this stage. Cycling frequency was added at the last step of model building and was found to be significant by Wald's test ($p = 0.013$). McFadden's ρ^2 values for the models with cycling frequency are ~0.7 (with base equally likely). Since the model fits were quite high, it was hypothesized that cycling frequency determines, to a large extent, the propensity of a cyclist to self-classify himself/herself into a particular category. However, at this stage model sample sizes were ~ 33% of the original sample sizes mainly because of missing data on income and cycling frequency. Since income was insignificant in all models, a final model was designed by removing income but leaving in cycling frequency which brought back the sample size ~ 50% of the original. The ρ^2 for this model was found to be slightly lower than the earlier model but in absence of income, age group 25-44 was found to gain significance. Age, gender, rider history and cycling frequency was found to have significant influence on whether a

cyclist classifies himself or herself into the categories of *strong and fearless* as well as *comfortable but cautious* and *interested but concerned*. The only significant predictor for the *enthused and confident* group was found to be cycling frequency and therefore, a model with only rider history and cycling frequency was built for this group and the ρ^2 was found to be ~ 0.6. Cycling frequency only model was found to provide a ρ^2 of 0.48 indicating that the propensity of cyclist classifying himself/herself into the *enthused and confident* category is well specified by his/her cycling frequency alone. It may therefore be suggested that cyclists who self-classify themselves into this category mostly do so because of their riding frequency rather than their self- perception on a confidence scale. As mentioned earlier, for all the categories, two final models are presented: one without cycling frequency and one with cycling frequency. Table 3a presents the model results for binary logistic models

The second variation on user's decision process was modeled using ordinal logistic models where the riders are thought of as classifying themselves into different categories *(y)* based on ordered partition of a latent continuous one dimensional confidence scale*(y*)* (*y = k, if* $\tau_{k-1} < y^* \leq \tau_k$ *where k = rider type categories in an ordered scale of 1 through 4, with 1 being least confident and 4 being most confident*). The model building exercise was the same as that for binary models and the results for the ordinal models are presented in Table 3b.

The ordinal logistic models are parsimonious and efficient as the choice is modeled on a single dimensional latent continuous variable. However, as mentioned by Bhat and Pulugurta (1997), it might be oversimplification of the actual decision process where the user is actually choosing among many alternatives the one alternative that he/she feels best satisfied with. In this case, the user has a *k*-dimensional choice space where *k* represents the number of choices faced by the user and estimating an unordered response using an ordered response model can lead to biases in

estimating probability of the choices (Bhat and Pulugurta 1997, Amemiya 1985). Therefore, the next set of models estimated were multinomial logistic regressions where the user was thought of as having to choose between the four rider type categories simultaneously ("Am I *strong and fearless* or *enthused and confident* or *comfortable but concerned*, etc."). The same model building exercise was followed in this case as with the binary logit models with the *comfortable but cautious* category treated as the base category. The first model included only age and gender and gave a McFadden's ρ^2 of 0.15. The final model, without income, included age group 18-24 and 25-44, gender, rider history and cycling frequency and gave a McFadden's ρ^2 of 0.6. The model with income and cycling frequency gave a model fit of 0.7 (unadjusted for model sample size). Age group 25-44 was found to be significant for the enthused and confident group when base group was changed to age group 18-24 indicating that cyclists in the age group of 25-44 behave significantly differently in self- classifying themselves into the enthused and confident group as compared to the age group 18-24. Chi-squared tests for model comparisons were not performed due to unequal sample sizes. Models with cycling frequency gave a higher McFadden's ρ^2 than the models without cycling frequency but were estimated on a much smaller sample size potentially removing a considerable amount of variation present in the dataset that was used for estimating the other models. Therefore, it cannot be definitively concluded that the models with cycling frequency are better models than their counterparts and hence, both types of models are presented in this paper. The multinomial logistic models are presented in Table 3c. Table 4 presents the odds ratios for the multinomial and the ordinal models both with and without cycling frequency.

Table 3a. Binary Logistic Regression Models

Co-efficients	Strong and Fearless Estimates (t-stat)		Enthused and Confident Estimates (t-stat)		Comfortable but Cautious & Interested, but concerned Estimates (t-stat)	
	Model 1 N= 740	*Model 2 N= 496*	*Model 1 N= 740*	*Model 2 N= 499*	*Model 1 N= 740*	*Model 2 N= 496*
Intercept	**0.239*** ** *(5.832)*	**0.329 *** ** *(6.846)*	**0.468 *** ** *(8.929)*	**0.531 *** ** *(11.406)*	**0.293 *** ** *(6.077)*	**0.167 ** ** *(2.832)*
Age			**Base: Age 45+**			
18-24	**0.102 .** *(1.763)*	0.0267 *(0.429)*	-0.082 *(-1.112)*		-0.02 *(-0.286)*	-0.05 *(-0.654)*
25-44	**0.11** ** *(2.986)*	**0.066 .** *(1.665)*	-0.002 *(-0.05)*		**-0.108 *** *(-2.475)*	**-0.1 *** (-1.978)
Gender			**Base: Male**			
Female	**-0.155*** ** *(-4.618)*	**-0.168 *** ** *(-4.452)*	0.01 *(0.243)*		**0.145 *** ** *(3.467)*	**0.158 *** ** *(3.428)*
Income			**Base: Income < $75,000**			
Income>= $75,000	0.007 *(0.248)*		-0.045 *(-1.142)*		0.037 *(1.03)*	
Rider history			**Base: Since Childhood**			
One year or less	**-0.235*** ** *(-5.779)*	**-0.169 *** ** *(-3.644)*	**-0.14 ** ** *(-2.685)*	-0.097 (-1.573)	**0.375 *** ** *(7.81)*	**0.269 *** ** *(4.733)*
Several years	**-0.143*** ** *(-4.512)*	**-0.135 *** ** *(-3.62)*	**0.081 *** *(1.982)*	0.063 *(1.258)*	**0.063 .** *(1.671)*	*0.069 (1.503)*
Cycling Frequency			**Base: Daily**			
Several times/week		**-0.071 .** *(-1.792)*		-0.08 *(-1.517)*		**0.144 ** ** *(3.0)*
Once or less/week		**-0.181 *** ** *(-3.88)*		**-0.183 ** ** *(-2.965)*		**0.357 *** ** *(6.252)*
Model Statistics						
Market Share of Group in the Model Dataset	*150 (20.27%)*	*90 (18.15%)*	*328 (44.32%)*	*226 (45.29%)*	*262 (35.44%)*	*180 (36.29%)*
Market Share of Other Groups in the Model Dataset	*590*	*406*	*412*	*273*	*478*	*316*
McFadden's ρ2 (Full model, base EL)	*0.368*	*0.630*	*0.460*	*0.636*	*0.473*	*0.660*
McFadden's ρ2 (MS model, base EL)	*0.177*	*0.177*	*0.292*	*0.292*	*0.251*	*0.251*
LL(0)	*-536.232*	*-536.232*	*-967.031*	*-967.031*	*-876.397*	*-876.397*
LL(MS)	*-441.357*	*-441.357*	*-684.42*	*-684.42*	*-656.533*	*-656.533*
LL(Full Model)	*-339.0786*	*-198.181*	*-522.546*	*-352.327*	*-462.058*	*-298.22*
G2=-[2(LL(Null)-LL(Full Model))]	*394.3068*	*676.102*	*888.97*	*1229.408*	*828.678*	*1156.354*

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 3b. Ordinal Logistic Regression Models

Co-efficients	Estimates (t-stat)	
	Model 1 *N= 740*	*Model 2* *N= 496*
Intercepts	**Base: Comfortable,but cautious & Interested, but concerned**	
Comfortable, but cautious & Interested, but concerned| Enthused and confident	**-0.938** *** *(-4.53)*	**-1.755** *** *(-6.357)*
Enthused and confident|Strong and Fearless	**1.26** *** *(6.005)*	**0.692** ** *(2.62)*
Age	**Base: Age 18-24**	
Age 18-24	0.316 *(1.065)*	0.151 *(0.451)*
Age 25-44	**0.622** ** *(3.357)*	**0.448** * *(2.072)*
Gender	**Base: Male**	
Female	**-0.823** *** *(-4.847)*	**-0.939** *** *(-4.546)*
Income	**Base: Income < $75,000**	
Income >= $75,000	-0.058 *(-0.385)*	
Rider History	**Base: Since Childhood**	
One year or less	**-1.791** *** *(-8.127)*	**-1.388** *** *(-5.209)*
Several years	**-0.554** ** *(-3.532)*	**-0.596** ** *(-2.994)*
Cycling Frequncy	**Base: Daily**	
Several times per week		**-0.638** ** *(-3.045)*
Several times per month		**-1.68** *** *(-6.361)*
Model Statistics		
McFadden's ρ2 (MS model, base EL)	*0.093*	*0.093*
McFadden's ρ2 (Full model, base EL)	*0.33*	*0.58*
LL(Null Model)	*-1086.527*	*-1086.527*
LL(MS Model)	*-985.081*	*-985.081*
LL(Full Model)	-723.306	*-459.281*
G2(Full Model, base EL)	*726.442*	*1254.492*
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*		

Table 3c. Multinomial Logistic Models

Co-efficients	Enthused and confident Estimates (t-stat)		Strong and fearless Estimates (t-stat)	
	Model 1 N= 740	*Model 2 N= 496*	*Model 1 N= 740*	*Model 2 N= 496*
Intercepts	**Base: Comfortable, but cautious & Interested, but concerned**			
Enthused and Confident	**0.51*** *(2.09)*	**1.183 **** *(3.546)*		
Strong and Fearless			-0.274 *(-0.84)*	**0.829 *** *(1.933)*
Age	**Base: Age 45+**			
Age 18-24	-0.122 *(-0.355)*	0.202 *(0.513)*	0.363 *(1.392)*	0.297 *(0.519)*
Age 25-44	**0.348 .** *(1.6)*	**0.394 .** *(1.576)*	**0.945 **** *(3.149)*	**0.731*** *(2.041)*
Gender	**Base: Male**			
Female	**-0.413*** *(-2.14)*	**-0.478 *** *(-2.072)*	**-1.64 ***** *(-4.833)*	**-2.199 ***** *(-4.39)*
Income	**Base: Income < $75,000**			
Income >= $75,000	-0.221 *(-1.19)*		-0.1 *(-0.421)*	
Rider History	**Base: Since Childhood**			
One year or less	**-1.305 ***** *(-5.38)*	**-1.053 **** *(-3.576)*	***-2.61 ***** *(-6.09)*	**-2.07 ***** *(-4.176)*
Several years	-0.061 *(-0.315)*	-0.156 *(-0.648)*	***-0.921**** *(-3.675)*	**-1.077 **** *(-3.135)*
Cycling Frequncy	**Base: Daily**			
Several times/week		**-0.771 **** *(-2.767)*		**-0.954 **** *(-2.789)*
Once or less/week		**-1.556 ***** *(-4.949)*		**-2.547 ***** *(-5.305)*
Model Statistics				
McFadden's ρ2 (MS model, base EL)	*0.09*			
McFadden's ρ2 (Full Model, base EL)	*0.34*	*0.58*		
LL(Null Model)	*-1080.56*	*-1080.56*		
LL(MS Model)	*-985.08*	*-985.08*		
LL(Full Model)	*-716.136*	*-452.958*		
G2(Full Model, base EL)	*537.89*	*1064.24*		

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 4. Odds Ratios for Multinomial and Ordinal Models with and without Cycling Frequency

	MNL Model 1		MNL Model 2		Ordinal Model 1	Ordinal Model 2
	Enthused and confident	Strong and fearless	Enthused and confident	Strong and fearless		
Age 18-24	0.884	1.89	1.223	1.345	1.372	1.163
Age 25-44	1.416	2.574	1.482	2.077	1.863	1.565
Gender Male	0.66	0.193	0.62	0.11	0.439	0.39
Income > = $75,000	0.8	0.904			0.943	
Rider history less than a year	0.27	0.0735	0.349	0.126	0.167	0.249
Rider history several years	0.94	0.398	0.859	0.341	0.575	0.551
Cycling frequency several times per week			0.462	0.385		0.529
Cycling frequency once or less per week			0.211	0.078		0.186

Overall, some distinct patterns were visible across all the models that we experimented with:

(1) Gender was significant in all the models with a negative sign, implying that female cyclists are more likely to classify themselves into low comfort/low confidence groups. The negative coefficients increase in value as we move from comfortable and cautious & interested, but concerned group to strong and fearless group which strengthens the previous inference. From the MNL models, being a female rider decreases the odds of being enthused and confident rider as compared to comfortable but cautious rider by more than 30% while the odds of

being a strong and fearless rider as compared to comfortable but cautious rider is decreased by about 80%

(2) Cyclists in the age group of 25-44 and 18-24 are more likely to be more confident riders than the cyclists in the age group of 45+. From the MNL model without cycling frequency, being cyclists in the age group of 25 to 44 increases the odds of classifying themselves into *strong and fearless* category by about 157% and of classifying themselves into enthused and confident category by about 42% more confident categories as compared to the cyclists in the age group of 45+ by while from the model with cycling frequency, being a rider in the age group of 25-44 increases the odds of classifying themselves into *strong and fearless* category higher confidence groups by about 100% and of classifying themselves into *enthused and confident* category by about 48% ; cyclists in the age group of 25-44 also have higher odds to classify themselves into higher confidence groups than cyclists in the age group of 18-24. This may be due to the inherent construct of the dataset where most users in the age group of 18-24 are students and use bicycle because they do not have access to car. Intuitively, they may be less bicycle enthusiast than riders in the age group of 25-44 who, being in the higher income group (also a construct of this dataset), may have access to automobile but still choose cycling as a mode of commute.

(3) Income is not significant but income greater than $75,000 is positively related to classifying oneself into *strong and fearless* and *comfortable but cautious & interested but concerned* group and is negatively related to classifying oneself into *enthused and confident* group.

(4) Riders with more experience may be hypothesized to be more confident as is captured by the negative coefficients of rider history of several years and rider history of one year or less as compared to the riders riding from childhood. For the MNL model without cycling frequency, the odds for riders in the several years category to classify themselves into the strong and fearless category decreases by ~ 60% and by 96% for the new riders. Similarly, as compared to the riders riding from childhood, the odds decrease by 6 % in classifying themselves as *enthused and confident* for the riders cycling for several years and by 73% for new riders. For the model with cycling frequency, the odds of classifying themselves into strong and fearless category decrease by 66% for the riders with several years riding experience and by 88% for new riders. The corresponding percentage decreases in the odds of classifying themselves into enthused and confident category are 15% for the riders with several years of experience and 65% for new riders, as compared to the riders riding from childhood.

(5) Cycling frequency is a significant determinant of rider type and higher frequency of cycling implies a more confident cyclist. Cyclists with cycling frequency several times per week and cycling frequency once or less per week are both less likely to be more confident than cyclists with cycling frequency daily. However, the magnitude of the co-efficient is higher in the once or less per week category than several times per week implying that cyclists in that category are even less likely than the cyclists in the several times per week category to be more confident riders. Cyclists who bicycle several times per week have about 62% lower odds to rate themselves into *strong and fearless* category and 54% lower

odds of classifying themselves into the *enthused and confident* category than riders who bicycle daily. Similarly, the odds for cyclists with cycling frequency once or less per week decrease by about 93% and by 80% to classify themselves into strong and fearless and enthused and confident category respectively as compared to daily cyclists.

(6) Since the ρ^2 are similar across binary, multinomial and ordinal models, it is difficult to justify the use of any one particular type of model for the purpose of cyclist classification. However, ordinal models impose an inherent restriction on the estimation process by assuming that the effect of the explanatory variables are the same at different category levels, i.e., how gender influences in self-classifying someone into a *comfortable but cautious* rider rather than an *enthused and confident* rider is the same as the influence of gender on being *enthused and confident* rather than *strong and fearless.* This may not hold true if the perceived difference in confidence between being *strong and fearless* and *enthused and confident* is smaller than the difference between *comfortable but cautious* and *enthused and confident*. Gender may have a much more pronounced effect on choosing whether a rider is *comfortable but cautious* as compared to *enthused and confident* than in choosing between *strong and fearless* and *enthused and confident* rider type. Therefore, conceptually, MNL models seem to be more appropriate for the purpose of this research.

Part 2: Understanding Infrastructure Preference Of Cyclists

In addition to the influence of socio-demographics on cyclist types, it is important to understand how cyclist type influences preferences for infrastructure. The basic

premise of this part of the research is to understand whether route preference is perception dependent and if that perception is a construct of the socio-demographic background of the cyclist. Based on model results presented in the first part of this paper, we hypothesized that female riders and riders in the age group above 44 years are more likely to prioritize safety over shortest route. Based on literature review (Hood et al. 2011, Sener et al. 2004), we also hypothesized that route impediments like high slope or poor pavement conditions are more likely to deter female cyclists from choosing that route.

Data Source: Cycle Atlanta User Survey and Atlanta Regional Commission Survey

The second part of the study is based on an online survey that was conducted among the users of the Cycle Atlanta application in the spring of 2014. The survey was sent to current application users via email addresses provided on the same user information screen asking demographics questions. The survey was divided into a few segments: (i) users' feedback on the Cycle Atlanta application, (ii) user feedback on the level of civic engagement and public participation in planning achieved through the Cycle Atlanta application, (iii) user feedback on which factors made them more likely to choose cycling as a mode of transport and (iv) socio-demographic and cycling related information of the respondents including self-classified rider type. The survey was sent to 697 application users and the particular question considered for this part of the analysis had 127 responses, a response rate of approximately 18%.

To increase the number of observations for this analysis, we appended the Atlanta Regional Commission's Bicycle User survey dataset to the existing Cycle Atlanta survey dataset. Both the Cycle Atlanta survey and the regional bicycle user survey were web

based surveys advertised through the same channels – the Cycle Atlanta survey was designed in accordance with the regional survey to preserve the comparability of the datasets. Since the regional survey went out to all the bicyclists in the Atlanta region, there is a possibility that it would include the Cycle Atlanta users as well and some respondents may be present in both the datasets. Euclidean distance matrices, which quantify the dissimilarity between rows of sample data were calculated individually for the Cycle Atlanta and the regional survey data and also after appending the datasets. The distance measures were found to be similar for both the cases and therefore, it was assumed that even though combining the two datasets may result in some data overlap, it will not significantly affect the results and the interpretation of the results. In addition, to maintain compatibility with Cycle Atlanta users, only the bicyclists in the regional survey having access to smartphones were included in the analysis. The non-smartphone respondents in the regional survey were instead used to verify that smartphone ownership does not bias results. Figure 8 shows the sociodemographic distribution of the pooled survey respondents across different rider types.

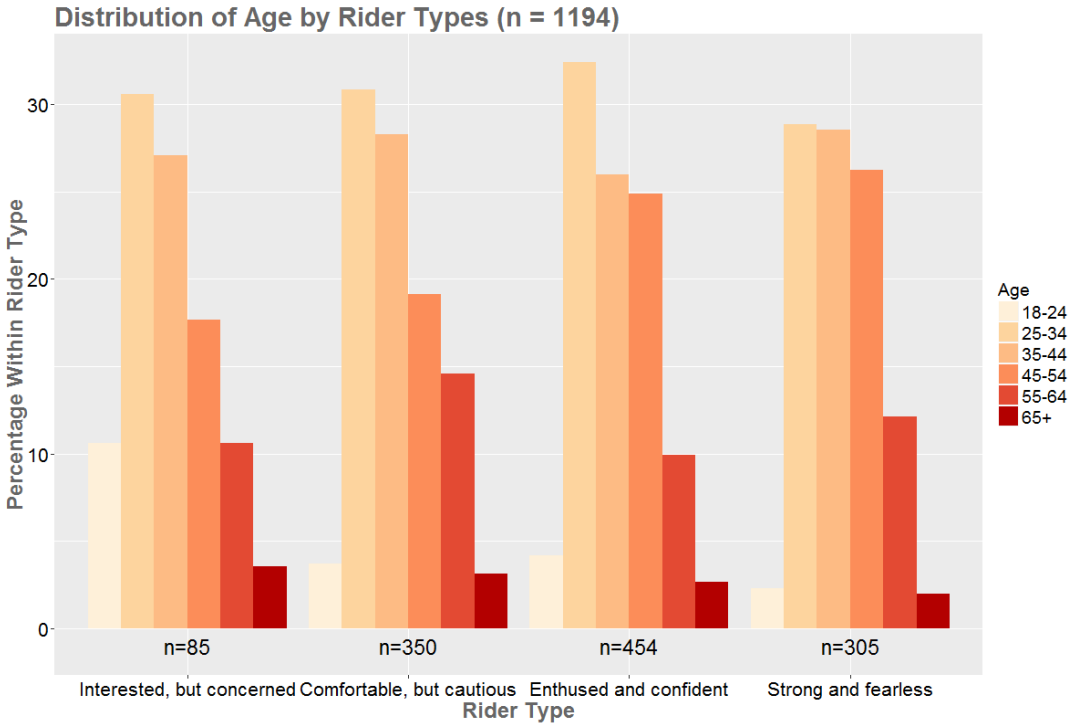


Figure 8(a). Distribution of Age by Rider Type (n = 1194)

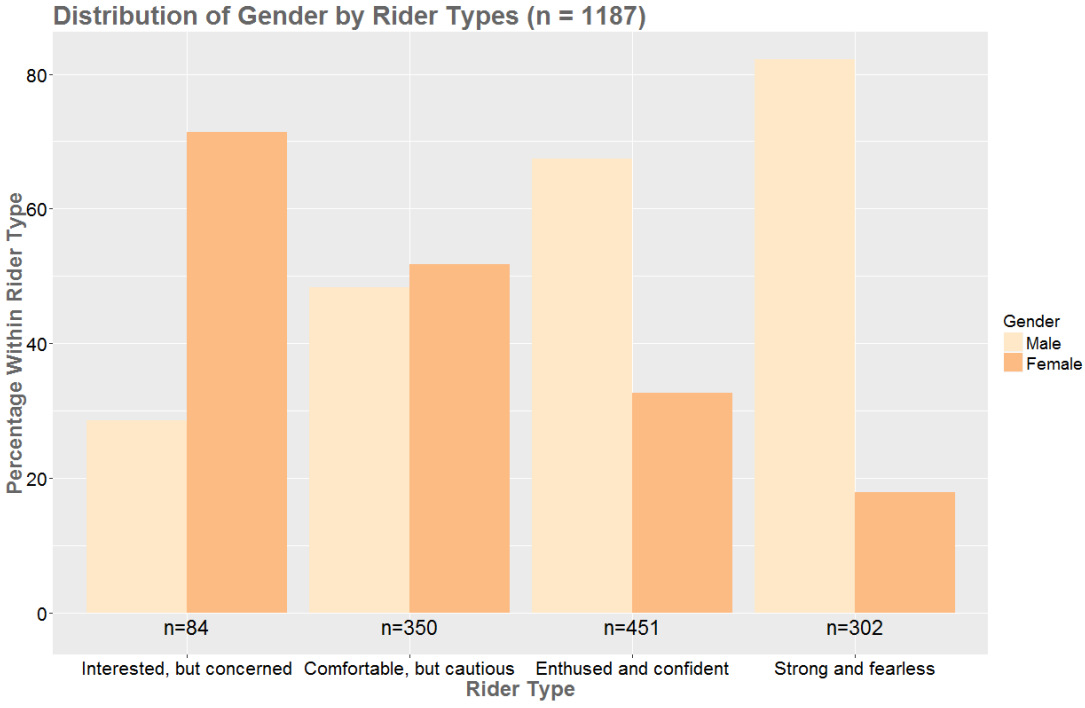


Figure 8(b). Distribution of Gender by Rider Type (n = 1187)

Figure 8 (a) – 8(c). Socio-demographic Distributions of Pooled Survey Respondents across Rider Types

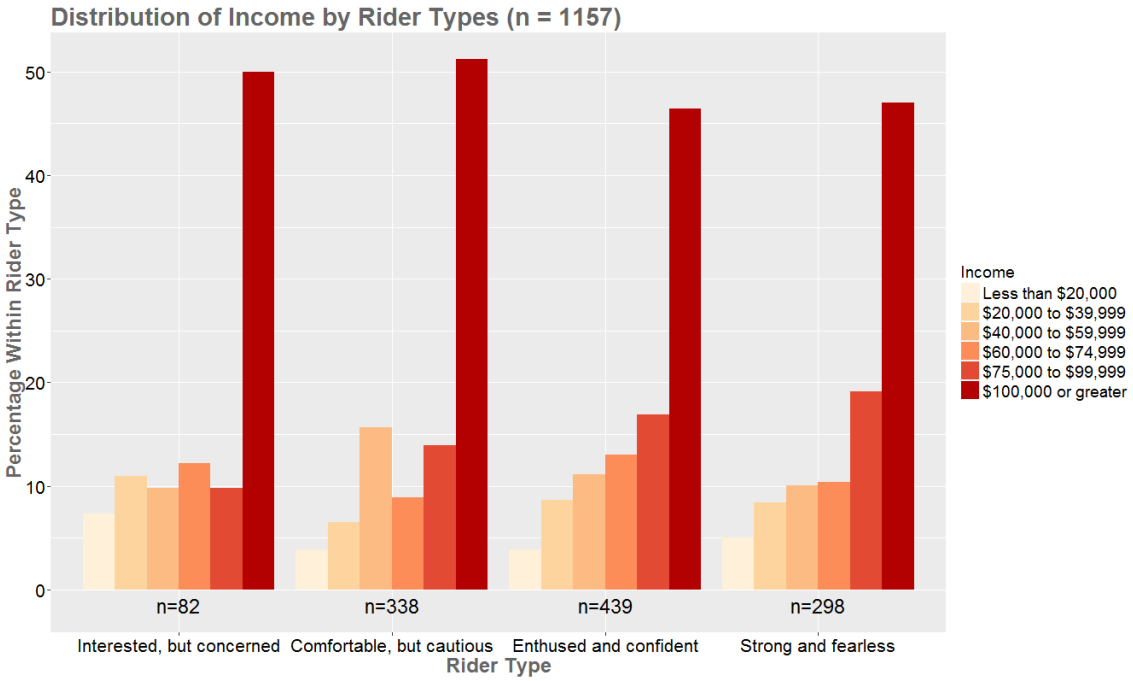


Figure 8(c). Distribution of Income by Rider Type (n = 1157)

Figure 8(a) – 8(c) Socio-demographic Distributions of Pooled Survey Respondents across Rider Types Continued

Table 5. Means and Standard Deviations of Item Responses on Road Conditions and Facilities by Rider Type

	Strong and Fearless		Enthused and Confident		Comfortable but Cautious		Interested but Concerned		Significance in difference in mean scores (ANOVA)
Conditions	Mean Score	Std. Deviation	Mean Score	Std. Deviation	Mean Score	Std. Deviation	Mean Score	Std. Deviation	
Bike Lane	3.13	*2.02*	3.41	*2.08*	3.47	*2.02*	3.29	*1.89*	***
Separate Path	3.00	*1.98*	3.39	*2.09*	3.65	*2.08*	3.72	*2.00*	***
Heavy Traffic	1.44	*1.35*	1.07	*1.25*	0.64	*1.10*	0.43	*0.98*	***
High Speed	1.03	*1.28*	0.60	*1.08*	0.47	*0.96*	0.31	*0.83*	***
Safe Routes	3.06	*2.08*	3.08	*2.12*	3.04	*2.10*	2.78	*2.10*	*
Directness	3.03	*1.96*	2.98	*1.93*	3.07	*1.95*	3.18	*2.03*	
Poor Pavement	1.57	*1.31*	1.42	*1.21*	1.20	*1.25*	1.11	*1.25*	***
Steep Hill	2.31	*1.43*	2.16	*1.41*	1.85	*1.40*	1.28	*1.22*	***
Parked car	2.23	*1.26*	2.18	*1.20*	1.97	*1.25*	1.68	*1.24*	***
Traffic Signal	2.15	*1.28*	2.20	*1.22*	2.00	*1.34*	1.79	*1.45*	*
Attractive Scenery	2.82	*1.83*	2.86	*1.76*	2.96	*1.78*	2.84	*1.79*	
Note:	*The responses were coded as a five point Likert scale(Much less likely =1, Much more likely = 5) Scores higher than 3.00 indicate a preference for that facility or condition while scores less than 3.00 indicate a negative impact of that facility or condition on choosing bicycling as an option. *** indicate significance at 0.001 level, ** indicate at 0.01 level and * indicate significance at 0.05 level*								

Table 6 p – values for Pairwise t-test on Respondents' Ratings on Influence of Road Conditions and Facilities on Bicycling Propensity, Paired by Rider Type

Road Condition	Rider Type			
Bike Lanes		SF	EC	CC
	EC	0.151		
	CC	**0.008**	0.048	
	IC	0.657	0.127	**0.000**
Separate Paths		SF	EC	CC
	EC	0.70421		
	CC	0.45595	0.552	
	IC	0.05538	**0.00033**	**1E-05**
Heavy Traffic		SF	EC	CC
	EC	0.16		
	CC	**1.00E-05**	**7.60E-07**	
	IC	**1.10E-08**	**1.20E-11**	0.014
High Speed		SF	EC	CC
	EC	0.1718		
	CC	**0.0037**	0.0121	
	IC	**2.30E-06**	**1.10E-07**	**0.0013**
Safe Route		SF	EC	CC
	EC	0.0439		
	CC	**0.0039**	0.1687	
	IC	**0.0154**	0.492	0.5504
Directness		SF	EC	CC
	EC	0.988		
	CC	0.214	0.036	
	IC	0.683	0.507	0.189

Road Condition	Rider Type			
Poor Pavement		SF	EC	CC
	EC	0.302		
	CC	**0.004**	**0.002**	
	IC	**0.003**	**0.002**	0.750
Steep Hills		SF	EC	CC
	EC	**0.00058**		
	CC	**7.40E-09**	**0.00014**	
	IC	**1.90E-09**	**3.50E-05**	0.47
Parked Cars		SF	EC	CC
	EC	0.02712		
	CC	**2.40E-05**	**1.02E-03**	
	IC	**1.10E-04**	**7.71E-03**	0.73144
Traffic Signals		SF	EC	CC
	EC	0.1818		
	CC	**0.0029**	**0.0074**	
	IC	0.0345	0.2100	0.2097
Attractive Scenary		SF	EC	CC
	EC	0.196		
	CC	0.016	0.072	
	IC	0.143	0.764	0.157

Note: Rider Type: SF = Strong and fearless, EC = Enthused and confident, CC = Comfortable, but cautious, IC = Interested, but concerned; Each cell value represents p-value for pairwise t-test between the row-column pair of rider type. Significant values are marked in bold.

Basic Statistics

In both the surveys, the respondents were asked to indicate on a five-point scale of less likely to highly likely, how the presence of route choice related factors may influence their decision to choose bicycling as a mode of transportation. The mean scores and standard deviations of each option were further calculated for each rider type and are shown in Table 5 along with the significance in difference of scores across rider types. Scores higher than 3.00 indicate a preference for that facility or condition while scores

less than 3.00 indicate a negative impact of that facility or condition on choosing bicycling as an option.

In general, bike lane and separate paths have a very high score across all rider types implying that all riders prefer dedicated facilities. Although this is generally not surprising, it is counter to the opinions expressed by some vehicular cycling enthusiasts (Forester 2012). Similar high scores are noted for safety and directness indicating that dedicated bike facilities along shortest routes are preferred by all riders. Negative conditions like poor pavement, steep hills, parked cars, and traffic signals negatively affect the decision to bicycle, but to a much lesser degree than traffic speed and volume. Traffic stress stands out as the most deterring factor preventing people to decide in favor of bicycling.

While average scores for individual conditions and facilities are mostly similar across the first three rider types, an ANOVA done on the scores show significant difference across rider types. To understand which groups actually differed significantly, a pairwise t-test was conducted. The p-values are shown in Table 6. For neighboring categories, there is no significant difference in mean scores across rider types strong and fearless and enthused and confident for any item while interested but concerned and comfortable but cautious groups have significant difference for traffic speed and separate facilities. There is significant difference across groups particularly for heavy traffic, high speed, poor pavements, steep hills and parked cars. There is no significant difference in item scores for directness of route and attractive scenery on route.

Further exploratory analysis was performed to validate the hypothesis that reported preference of infrastructure and facilities depends on the socio-demographic

attribute of the users. First, a factor analysis was performed to group the 10 road conditions and facility preferences into fewer factors. Then, regression analyses were done with each factor as the dependent variable and the sociodemographic attributes of the respondents as the explanatory variables.

Factor Analysis

Factor analysis is used to address underlying correlation among all or some of the observed variables such that there will be multi-collinearity issues if the variables are treated individually in the data (Thompson 2004, Hurley et al. 1997). Factor analysis, thus, helps in reducing the dimensionality of the dataset as well as in identifying the correlated structure of observed variables. For example, among the 10 variables influencing the decision to bicycle, high traffic speed and heavy traffic volume may be correlated (that is people who are averse to high traffic speed are also likely to be averse to heavy traffic volume). By using factor analysis, we may be able to group these two variables together to form a new factor variable which can then be used for regression. This will reduce the number of variables from 10 to 9 and will remove the collinearity that would occur if both the variables are treated separately in regression.

Factor analysis can be exploratory or confirmatory, the latter being used to test a pre- determined hypothesized correlation among some of the variables (Hurley et al. 1997). In our case, no correlation structure was initially hypothesized, and hence, an exploratory factor analysis was performed. The variables were allowed to load into all factors, irrespective of their score, and an orthogonal rotation was used. A scree plot was used to determine the optimum number of components; and two models, one with two

factors and another with three factors, were tested. The three factor model was used for further analysis. Table 6 presents the results of the three factor model.

Table 7. Exploratory Factor Analysis: Loadings

Loadings:	Factor1	Factor2	Factor3
Bike lanes	0.884	0.337	
Separate paths	0.89	0.316	
Safe route	0.767	0.389	
Directness	0.797	0.357	
Attractive scenary	0.76	0.405	
Steep hills	0.445	0.577	
Parked cars	0.43	0.714	
Traffic signals	0.443	0.645	
Heavy traffic			0.778
High traffic speed			0.782
Poor pavement	0.399	0.42	0.386
	Factor1	**Factor2**	**Factor3**
SS loadings	4.173	2.164	1.763
Proportion Var	0.379	0.197	0.16
Cumulative Var	0.379	0.576	0.736

Table 7 shows that factor 1 has high loading on bike lanes, separate paths, safe route, directness, and attractive scenery. This factor was named Protected Environment as the preference of people scoring high on this factor appears to be direct and safe facilities. The second factor has moderately high scores on steep hills, parked cars and traffic signals and was therefore named Route Impedance implying that people who score high on this factor prefer routes with less disruption. The third factor has the highest loading on heavy traffic and high traffic speed and was named Route Stress indicating that people who score high on this factor are averse to traffic stress: their decision to bicycle is largely determined by the traffic speed and volume in the corridor.

Regression Analysis

In the second stage, to understand if sociodemographic attributes of riders influence the infrastructure preference, the three factors were used in regression equations

with age, gender, income, and rider type as explanatory variables and the factors as the dependent variables. Table 8 provide the details of the regression analysis.

For the factor Protected Environment, gender, income, and rider type are significant implying that females and those in high income groups prefer facilities. The rider type variable has a negative coefficient indicating that people with lower confidence levels prefer separate facilities. Both Route Impedance and Route Stress are factors with negative connotations and the regression results should be interpreted accordingly. Age, income, and rider type are significant for the factor Route Impedance, which includes steep hills, parked cars, and delayed traffic signals. Age shows a negative sign implying that older riders have a stronger aversion to route impedances like steep hills or parked cars, while rider type shows a positive coefficient implying that less confident riders are more averse to route impedance.

For the Route Stress factor, age, gender, income, and rider type are significant variables. Age, gender, and income have negative coefficients while rider types has a positive coefficient. The result can be interpreted as older and female riders, as well as riders in the high income group, are less likely to decide to bike under traffic stress while riders in the low confidence category are also deterred from bicycling because of high traffic speed and volume.

It should however be noted that all the regression models have low R-squared values, ranging from 0.04 (Protected Environment and Route Impedance) to 0.1(Route Stress). Therefore, while it can be implied that infrastructure preference affects the decision to bicycle differently for older, female, and less confident riders, it is also clear

that there are other factors that influence the decision to bicycle which are not included in these models.

Table 8. Regression Analysis for Protected Environment, Route Impedance and Route Stress

	Protected Environment				Route Impedance				Route Stress			
Coefficients	**Estimate** *(t-stat)*	**Sig.**	**Estimate** *(t-stat)*	**Sig.**	**Estimate** *(t-stat)*	**Sig.**	**Estimate** *(t-stat)*	**Sig.**	**Estimate** *(t-stat)*	**Sig.**	**Estimate** *(t-stat)*	**Sig.**
(Intercept)	**-0.304** *(-2.304)*	*	**-0.964** *(-6.138)*	***	**0.48** *(3.702)*	**	**-0.231** *(-1.488)*	.	**0.172** *(1.595)*	.	**-0.501** *(-3.888)*	**
Age	**-0.106** *(-4.372)*	***	**-0.085** *(-3.491)*	**	**-0.082** *(-3.329)*	**	**-0.066** *(-2.721)*	**	**-0.047** *(-2.282)*	*	**-0.029** *(-1.45)*	.
Gender	**0.157** *(2.833)*	**	**0.187** *(3.273)*	**	**-0.197** *(-3.547)*	**	-0.042 *(-0.744)*		**-0.077** *(-1.672)*	.	0.05 *(1.059)*	
Income	**0.0723** *(3.791)*	**	**0.062** *(3.323)*	**	-0.002 *(-0.1)*		0.003 *(0.154)*		0.010 *(0.658)*		0.013 *(0.84)*	
Strong and fearless			**0.443** *(4.735)*	***			**0.783** *(8.482)*	***			**0.677** *(8.829)*	***
Enthused and confident			**0.821** *(9.424)*	***			**0.476** *(5.527)*	***			**0.53** *(7.418)*	***
Comfortable but cautious			**0.8** *(8.994)*	***			**0.28** *(3.189)*	***			**0.336** *(4.602)*	***
Model Statistics	Multiple R-squared: 0.030, Adjusted R-squared: 0.028 F-statistic: 12.2 on 3 and 1181 DF, p-value: 7.261e-08		Multiple R-squared: 0.116, Adjusted R-squared: 0.111 F-statistic: 25.48 on 6 and 1167 DF, p-value: < 2.2e-16		Multiple R-squared: 0.019, Adjusted R-squared: 0.016 F-statistic: 7.552 on 3 and 1170 DF, p-value: 5.26e-05		Multiple R-squared: 0.084, Adjusted R-squared: 0.08 F-statistic: 17.93 on 6 and 1167 DF, p-value: < 2.2e-16		Multiple R-squared: 0.006, Adjusted R-squared: 0.004 F-statistic: 2.503 on 3 and 1170 DF, p-value: 0.0579		Multiple R-squared: 0.076, Adjusted R-squared: 0.071 F-statistic: 16 on 6 and 1167 DF, p-value: < 2.2e-16	

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Discussion of Results

A substantial component of the self-classification of rider types may be correlated with the socio-demographic make-up of a cyclist. In particular, gender and age have a demonstrated effect on an individual's attitude towards safety, comfort, and confidence. Studies on the effect of gender on confidence have shown that females are much less likely to undertake risky tasks and more likely to report themselves to be less confident than their male counterparts even when performing identical tasks (Kray et al. 2001). A study by Byrnes et al. (1998) showed that this gender gap increases with increasing age

although the level of gap is decreasing over time. Emond et al. (2009) also emphasized the difference in individual and social perceptions between the two genders and its strong influence on self-efficacy.

This study was undertaken to provide a data driven answer to the question whether cycling infrastructure preference depends on the socio-demographic characteristics of the cyclists. We also aimed to validate the generalizability of the existing popular classification of cyclists based on their comfort and confidence level via information provided by the cyclists themselves. Results of our first analysis show that age, gender, and rider history influence self-classification of the riders into rider types. Overall, the confidence and comfort level decreases with age and is significantly lower for female riders as compared to their male counterparts. Cycling frequency and riding history both have a significant role in determining rider type but are positive reinforcement only through an adoption phase. The binary logistic regression model fits are comparatively lower than the ordinal logistic and ordered probit models indicating the validity of an underlying scale of confidence and comfort in the construct of rider type categories. Future research should aim to validate the claim based on revealed preference route choice data.

The purpose of the second part of the analysis was to understand if perceptions about route characteristics and safety are related to a cyclist's sociodemographic make up and self-perception as a particular rider type from a stated preference dataset and without any input from revealed route choice preferences. The results indicate a preference of dedicated facilities across all rider types and also point toward high speed traffic and high volume traffic as the factors that most negatively affect willingness to cycle. Further

analysis reveals that female riders are more likely to prefer separate facilities while elderly riders are more likely to be averse to steep hills and other route impedance. Less confident riders prefer both separate facility and less impedance in their routes. Higher income riders prefer protected environments but are not as deterred by presence of slopes or traffic signals on their routes. High traffic speeds and heavy traffic volumes deter females, older riders, and less confident riders from bicycling.

Limitations

The innovative use of smartphone based application to collect revealed preference cyclist route choice data has its own caveats. The Cycle Atlanta data suffer from the issues of self-selection bias as it is a crowd sourced data collection system and people who participate are those who are sufficiently interested in the project, willing to share data, and invest time without any personal gain. As a result, the Cycle Atlanta dataset is heavily dominated by male white cyclists in the age group between 25-44 years. This is not a representative sample of the population of Atlanta where 50% of the population is female and about 54% is African American (ACS 2012). However, there is currently no reliable estimate on the makeup of the cycling population of Atlanta (Poznanski 2013), and hence it is difficult to comment on the representativeness of the Cycle Atlanta data with regard to the cycling population of Atlanta. We compared the sociodemographic distribution of the Cycle Atlanta users to the participants of the Atlanta Regional Commission data and found no statistically significant difference. While this may mean that the cycling population of Atlanta is fairly homogenous, it may also be due to the reason that Atlanta Regional Commission's survey was advertised through pro bicycle channels that typically reach out to people with similar attitudinal preferences as the

Cycle Atlanta users. It is an ongoing future research debate as to whether weighting the data by Atlanta population proportion should be considered as that may interfere with the representativeness of the collected data.

The other issue associated with the data is that by design, the data systematically misses the cyclists who do not own a smartphone. A study on smartphone ownership (Windmiller et al. 2014) has shown that this systemic bias affects people in older age groups, certain ethnic groups, and sometimes people in lower income groups. The Cycle Atlanta dataset is sparse in all of these categories – the users are mostly White and high income in the age group of 25- 44. It is difficult to estimate how much of the sparsity is caused by use of smartphone for data collection and how much is due to characteristics that define cyclists of Atlanta. However, the number of non-smartphone owning Atlanta Regional Commission participants and Cycle Atlanta participants were similar in number, therefore making up a similar portion of the sample. The models were run for the non-smartphone owners as a separate group and no statistically significant difference in infrastructure preference was noticed. Therefore, in spite of the biases in the data collected via smartphone app, any infrastructure requirement predicted based on Cycle Atlanta data can be assumed to hold true for non-smartphone users as well.

Conclusion And Future Research

This research provided an analytical approach to understand the characteristics and preferences of the different types of cyclists. Cyclists using the Cycle Atlanta tracking application were given the option to self-classify themselves into rider types, and socio-economic data were collected to understand the basis for riders' choice of rider

type. The rider type classification was also used in a stated preference survey on route choice attributes to understand if such preferences are influenced by rider types.

The first part of the analysis shows that socio-demographic variables and riding pattern are significant predictors of a cyclist's probability of self-classifying himself/herself into a particular category. In particular, gender, rider history, and cycling frequency are significant in all the models. The results indicate that knowing a cyclist's demographic information can potentially help in classifying the cyclist into a particular rider type. In the future, this can help us to streamline surveys by replacing sociodemographic questions by a single rider type classification question. Alternatively, by knowing the socio-demographics characteristics commonly available through census data and other surveys, we will also be able to predict the rider type and hence infrastructure preferences of people without having to undertake a new survey design for cyclists only. It will also help in understanding infrastructure and facility need of future cyclists who are not yet cycling and hence, there are no revealed preference data on the preference of such future cyclists currently.

The results also direct attention to the requirement of segmented route and facility preference decision models for different cyclist types. Since the purpose of the route and facility preference analysis is to understand the requirements by rider types, segmented models based on rider type may enable a planner to better predict the choices of a future cyclist based solely on the demographic information of the cyclists. Future route decision model research may therefore explore segmentation of the dataset to achieve better predictability.

From the second part of the analysis, it is evident that most route perception issues and facilities are viewed on a similar scale by cyclists as the mean scores on those facilities are quite similar across rider types. Other results indicate that sociodemographic attributes and confidence levels influence infrastructure and facility preference. However, the model fits are substantially low indicating that rider level data are not sufficient to predict route level decision process. Further investigation is necessary, as the literature shows that choice of route depends on route characteristics as well as rider characteristics like age and gender. Therefore, in future, for further insight, we plan to augment user data by revealed preference route choice data to make any definitive conclusion about the preference and requirements of cyclists.

CHAPTER 4

DATA CLEANING AND MAP MATCHING

Introduction

Traditional travel behavior related data collection methods depend on surveys where the respondent is either required to recall an experience or incident or extrapolate present day experience to a future scenario. In both cases, the collected data suffers from respondents' personal bias and is dependent on recall efficiency of the respondents. In particular, surveys that span for multiple days and record multiple events are more likely to suffer from missing data on events that the respondent did not consider important for the reporting purpose. For example, household travel diary surveys that are generally used to record all household trips over a week typically tend to under report short trips and connection trips (walk to transit, errand trip during lunch). The other problem with multi - day surveys is that respondents suffer from survey fatigue towards the end of the survey period, leading to high rates of attrition and often, a low response rate to begin with.

Replacing or complementing traditional survey methods with data collected via advanced technologies like sensors and GPS enabled devices is gradually becoming popular among transportation researchers across the world (Shen and Stopher 2014, Du and Aultman-Hall 2007, NCHRP 2014). The ability of such technologies to record data without direct effort of the participant as well as the ability to capture revealed preference data of the participants in real time makes these technologies immensely usable for intelligent transportation systems where routing decisions are dynamic and made in real time. As recall effort is removed, the data quality is also improved substantially even

though data collected through GPS is not always completely accurate. Multiple studies have shown that GPS devices capture 20% – 30% more trips than traditional survey methods (Bricka & Bhat, 2006; Stopher & Greaves, 2009; Stopher & Shen, 2011) The use of GPS for data collection becomes particularly useful for user groups that are small in number but have distinct trip characteristics and hence infrastructure preferences such as pedestrians or bicyclists. Planning for such groups without data from the users themselves can be seriously flawed and has proven to be significantly less effective for encouraging people to use non-motorized transport. The regional planning agencies, however, are often unable to allocate sufficient funds and human resources to conduct a separate data collection effort for users like cyclists who constitute only about 1 % of all transportation system users. Towards that end, there has been a shift recently towards designing participatory planning processes that help people come together on a virtual platform or to provide indirect input to the planning process through data contribution. These methods rely on people's willingness to participate in the planning process and to share data voluntarily, and passive data collection technologies play a significant role in facilitating such platforms. These platforms, if successful, can remove additional data collection burdens for the planning agencies and can bridge the gap between data need and data availability. Cycle Atlanta is one such platform which was created for cyclists to voluntarily come together to better inform the City of Atlanta of their preferences and requirements.

The primary purpose of the Cycle Atlanta app is trip recording, in which the app uses the GPS capability of the phone to record the location of a user on a second by second basis and the trip is uploaded when it is complete. This offers a high level of

precision for recording trips over time, allowing for detailed research on route preferences not previously available on mass-scale data. However, to enable studies involving route choice and decision making models, these GPS points must be processed so that a trip follows the existing system of linear paths which comprise the road network. This task is non-trivial, particularly for the scale of Cycle Atlanta, in which we are analyzing about 20,000 trips with each trip consisting of about 1000 geocoded points. In this chapter, we present the preprocessing routines as well as two map matching procedures that were designed and used to process the Cycle Atlanta data. Parts of the preprocessing routines were borrowed from existing literature while parts were designed in house and proved to be very effective in reducing the computational burden at later stages of analysis. The map matching algorithms were developed as a combined outcome of a decision process on the part of the user as well as available network characteristics. The entire process of cleaning the GPS data and matching trips to network was carried out using open source software R and the code will be freely available in the GitHub repository of Cycle Atlanta.

Literature Review and State of the Practice

This research is based off data collected via GPS enabled smartphones and hence, the literature review presented here is focused on issues that are commonly found in the data collected through GPS. It should be noted though that there are other passive data collection devices like stationary sensors and in-vehicle information systems which have their own advantages and disadvantages that are beyond the scope of this paper and hence not covered here.

GPS enabled devices are capable of recording the latitude, longitude, heading, speed, altitude, and timestamp of a user at intervals of one second, three seconds or five seconds. Generally, data is collected per second but Shen and Stopher (2009) compared data adequacy for trip identification and mode detection across different time intervals and concluded that a five second interval provides sufficient data for trip and mode identification while reducing the number of GPS points to be dealt with. The location point is recorded via triangulation of signals from at least 3-4 satellites and issues with GPS data are mainly related to loss of signals or reflection of signals across high rise buildings.

There are two different types of GPS devices that are used for data collection – (1) GPS units either connected to a hand held Personal Digital/Data Assistant (PDA) or carried separately by the participant – in both cases, the GPS unit is dedicated to the purpose of data collection and it is the respondents' responsibility to monitor the unit and, often, record additional data such as sociodemographic data that the GPS unit cannot capture (Du and Aultman-Hall 2003, NCHRP 2014) and (2) GPS enabled devices that passively collect location data via some application – the primary purpose of the devices is not collection of location data but having GPS capability enables it to record location data of the user with timestamps. The respondent burden is hugely reduced with these devices as the users do not have to take part directly in the data collection system.
The data collected via passive GPS enabled devices often suffer from noise and uncertainty that require a substantial amount of post processing efforts to render the data usable (Shui and Shalaby 2007, Quddus et al. 2007, Pyo 2009, Auxhaussan 2012). The inaccuracy occurs from multiple sources - loss of signals at certain locations, particularly

at the start and end of the trip (cold start/warm start), reflection of signals between tall buildings before it reaches the device (urban canyon effect), not having enough satellites for accurate triangulation, and interference of signals at intersections. The data cleaning methods developed to deal with these problems are generally rule-based and use the number of satellites, speed and heading change, as well as position jump to identify points that are part of the trip (Stopher et al. 2005, Lawson et al. 2010, Shen and Stopher 2014, Wolf et al. 2001).

Trip Start and End

One of the most difficult parts of GPS data cleaning is identifying trip start and trip end points. At the start of the trip, the GPS device requires some time to acquire satellite signal, stabilize and then start recording the trip. Until that time, the signal jumps, resulting in a scattered cluster of points which make it difficult to identify the actual start or end point of the trip. Since the GPS records data at a regular interval, the most intuitive approach to identify trip end points is to identify points where the time interval between any two consecutive points is more than a specified dwell time. Multiple studies have used this to identify trip start and end points, albeit with different dwell time criteria. The most commonly used time gap is 120 seconds, the Highway Capacity Manual prescribed maximum signal timing, so that trips are not terminated when they are stopped at the signals (Schonfelder et al . 2002, 2003, Wolf 2001, Du and Aultman-Hall 2007, Stopher et al. 2005). Other studies have used dwell time from 45 seconds (Pearson 2001) to 3 minutes (Doherty et al. 2000). However, in most cases, dwell time based identification is supplemented by other criteria like zero speed, zero change in bearing (Doherty et al .2000, Schussler and Axhausen 2009, Lawson et al. 2010), difference in

latitude and longitude (Stopher et al. 2005) and point density (Schussler and Axhausen 2009). Since most of the methods were developed for vehicular traffic, Schussler and Axhausen 2009, Doherty et al. 2000, and Stopher et al. 2003 used engine stop and start time difference as a measure of trip end and start identification too.

Du and Aultman-Hall (2005) used a maximum and minimum dwell time criterion along with distance from network and bearing changes to identify trip ends. They used a buffer distance of 15 meters from road centerline and GPS points outside this buffer were discarded. For the points within that 15-meter buffer zone, dwell time and bearing changes were used to identify trip start and end points. Du and Aultman-Hall (2005) experimented with multiple combinations of minimum and maximum dwell times and the algorithm was tested for minimum dwell times of 20, 40 and 60 seconds while maximum dwell times tested were 60, 100 and 140 seconds. They found that any dwell time between the maximum and the minimum is generally associated with a 180^{o} bearing change.

Besides identifying trip start and end points, there are also issues with signal loss and signal noise. NCHRP 775 provides a comparison of three methods for dealing with GPS noise filtering (Stopher et al.2005, Schussler and Axhausen 2009, Lawson et al. 2010) and compares the results with a base case where the actual trip is known. All the three methods use dwell time threshold, number of available satellites and a threshold value of horizontal dilution of precision for noise filtering, along with zero speed and zero heading change. The report classifies the error of not removing an invalid point that has been removed in the base case as a Type 1 error while removing a point not removed in the base case as a Type 2 error. The analysis of three major data cleaning methods

show that all the methods tend to have more type 1 errors than type 2 errors, whereby the methods tend to retain more points than making the error of removing a point that is part of the trip (NCHRP 2014).

Map Matching

Map matching is the process of relating input data from global positioning systems to a spatial road network map to correctly identify the position of a vehicle on the road network (Quddus et al. 2007, Zhou and Golledge 2006, Axhausan et al. 2009). For transportation related studies, as Zhou and Golledge (2006) mention, the map matching process is a means of transferring road attributes to the travelled route so that further inferences can be made about travel patterns and preferences.

Map matching can be either a real time process which is used for most ITS related applications or can be done as a post processing step where the vehicle tracking information is not essential to update network information instantaneously. For this research, map matching was done as a post processing step of data collection and cleaning. The process of map matching can be either a point to point matching, point to line matching, or curve to curve (polyline to polyline). The process of matching is significantly complicated and is highly prone to errors due to the compounded effect of uncertainty of GPS points as well as inaccuracy in the road network. Quddus et al. (2007) mentions the accuracy of the match depends both on the quality of the road network map as well as the algorithm used because different algorithms may provide different efficiency for the same map.

Map matching algorithms are primarily classified as geometric, topological, and probabilistic. Other advanced methods that have been used for map matching include

adaptive fuzzy logic (Kim and Kim 2001) and Bayesian belief theory (Zhou and Golledge 2006). The geometric algorithms generally only take into account the distance of the GPS point to the road segments and return the nearest segment to the GPS point as the matched segment. The biggest issue with distance based matching is that if there are parallel segments that are sufficiently close to each other, then the GPS point can match to the wrong link. This is particularly true if the correct link has less nodes than the wrong link or if the matching buffer zones overlap. To overcome these issues, other parameters can be added to the matching criteria which include travel direction and bearing change, road attributes like one way lanes, speed limits and distances travelled on a segment (Najjar and Bonnifait 2003, Taylor 2001). In view of completing multiple parameter matching criteria, Quddus et al. (2003) suggested using a weighting approach to decide on the best match while Kim and Kim (2001) proposed an adaptive fuzzy network based training for the same purpose.

The issue with not using network information for the matching process often can result in matches that do not form a logical path which is important for transportation related purposes. The topological approaches to map matching restrict the matching using topological properties of the underlying network like link connectivity and altitude difference (Greenfield 2002, Ochieng et al. 2003). However, constructing a path from link connectivity depends on successive link matches which can result in a wrong path even if one of the links is matched wrong. To address this issue, Pyo et al. (2001) and Marchal et al. (2004) suggested keeping multiple candidate solutions for each GPS point along with some measure of goodness of fit for each candidate. At the end of the

matching process, the measures are aggregated and a match is decided based on that score.

Statistical map matching methods have used linear regression models to fit the GPS points to a road network (Lakakis 2000). Bierlaire et al. (2013) proposed a map matching algorithm that first generates a path between origin and destination points and then calculates the likelihood that the GPS points are generated along that path using geographical and temporal information. However, since the path is generated based on shortest distance between the origin and destination points, if the vehicle uses any other criteria for path choice, it will be difficult to find an appropriate match.

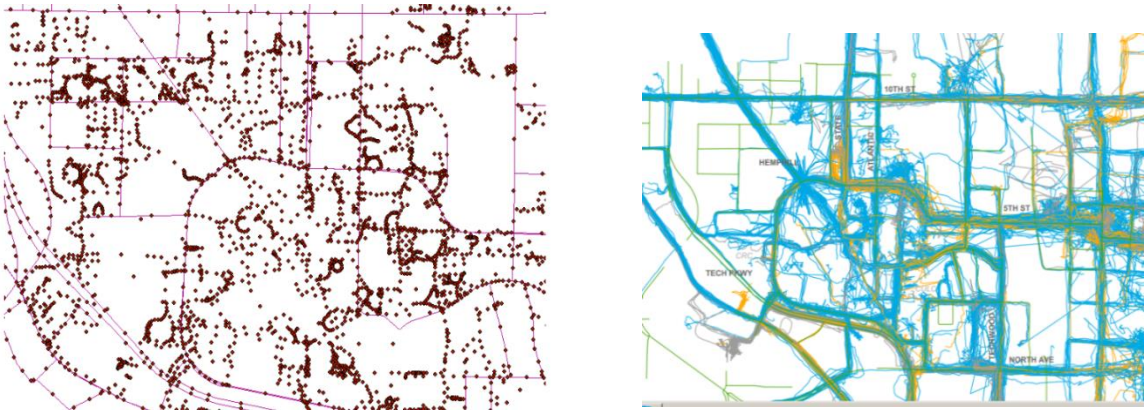
Methodology

In this section, we discuss the methods we used to collect, prepare and match the GPS data to the road network map of Atlanta. Multiple platforms and algorithms are available for most data processing steps and the related ones were tried and tested within this research. However, because bicycling trips tend to differ from vehicular trips, using algorithmic approaches developed for car traffic proved to be not very effective for either cleaning or matching. For example, since bicyclists often trade off shortest path for safest path or other considerations like scenery, any map matching algorithm based on the shortest path approach could not be used. Some algorithms could not be used because of the scale of application – this research used 15 million GPS points and a road network with more than 18,000 links. In addition, bicyclists are likely to have more options at each stage of the trip and may not be required to follow the conventional routes of vehicular traffic, making it more ideal to be modeled as a decision process at every intersection than a predetermined route between origin and final destination. Therefore,

most of the methods used and described here were designed keeping in mind the particular nature of the dataset and were modified as required iteratively during the process.

Data Collection

The GPS data used for this study were collected via the smartphone application Cycle Atlanta. Launch of the app in October 2012 was announced by the Mayor of the City of Atlanta and the app was widely publicized through various cycling advocacy groups and social media. Participation in using the app is voluntary and no reward was offered to record trips. The app is designed for both Android and iPhone GPS-enabled smartphones and is freely available for download from the app stores. The user has to turn on the app at the start of the trip and geolocation of the user is recorded from that point until the user indicates a trip end. The trip is not saved unless the person wants to save the trip which s/he can indicate via the 'save' button. At that point, the trip is uploaded to the secured database maintained by Georgia Tech. For each trip, the app records latitude, longitude, altitude, speed, time, and horizontal and vertical accuracy at an interval of 1 second. Figure 9(a) and 9(b) shows an example of the original uncleaned data from the Cycle Atlanta app.



(a) (b)

Figure 9. Original Uncleaned Data: (a) Raw GPS Points (b) Trip Lines Constructed from GPS points

Data Cleaning and Noise Filtering

The data issues found were classified as (1) systemic, (2) operational and (3) random. Systemic errors include issues that occur because of the use of GPS capability and are general in nature across all studies using GPS data. For example, cold and warm start problems, signal loss issues and urban canyon effects will be classified as systemic errors within our classification system. Operational errors are often errors introduced in the system by the users. These issues include forgetting to turn off the app after trip completion, using the app for non- cycling trips, using shortcuts and bylanes that are not part of the street network, etc. These errors will depend on the purpose of data collection and consequently on the participants. Random errors are most often related to systemic errors brought into the data due to use of GPS, but the nature of the errors are specific to each instant of recording and, hence, no standardized method can be applied to remove such errors.

The data cleaning was done following established practices from the literature. However, knowing the difficulty of map matching with noisy data, a lot of effort was put

into early cleaning stages before running the snapping algorithms and therefore, the standard practices were modified and customized to suit our needs. Some additional criteria were also implemented keeping in mind the specific nature of the dataset. Efforts were made to attain a balance between retaining as much necessary information as possible in contrast to retaining data that is erroneous and can increase the computational burden for a later stage of analysis. It should also be noted that the app did not report the number of satellites, so that information could not be used for data processing in our case.

Operational Error Handling

As the study focused on bicyclists in Atlanta, at first, the data were checked for geographical limits – since the app is freely available to anyone owning a smartphone, it was suspected that the data might have trips that are not Atlanta based. Therefore, any point with latitude and longitude beyond the latitude and longitudinal boundaries of Atlanta [NW: 33.886823, -84.551068; SE: 33.647808, -84.28956] was removed from the dataset. Some trips were recorded over multiple days which can happen if the user forgets to turn off the app at the end of a trip and the app continues to record trips as continuation of the first trip until it is turned off. In such cases, the day with maximum number of recorded points was retained and data from other days were discarded.

Random Error Handling

Duplicate removal and basic data filtering: Two types of duplicates were identified: (1) points within same trip having same timestamp but different latitude and longitude and (2) identical latitude, longitude, timestamp and user id but different trip id. So, while in the first case, all points except the first point are removed, in the second case, the trip with the lower trip id is retained and the duplicates are removed. Some

points were recorded with invalid timestamp (0000-00-00, 00:00:00) – these points were also removed during this step.

Horizontal Accuracy: As mentioned in NCHRP report and used in other research, the horizontal accuracy (haccuracy) threshold could be between 5 and 20 for a point to be a valid point. For this research, haccuracy limit was set to 30 – any point with horizontal accuracy more than 30 was removed from the database. The higher-than-standard limit was set after experimenting with haccuarcy values of 10, 20 and 30. Since the data are from cyclists who tend to use bylanes, cut throughs and underpasses which do not always have a good signal, setting a higher accuracy expectation resulted in removing too many points and created connectivity issues as well as sparse data problem for shorter trips.

Systemic Error Handling

Speed, Distance and Heading: The app recorded instantaneous speed at each point as well as latitude and longitude. Since the app is designed for cyclists, points with instantaneous speed more than 12 mph were discarded. Points with zero speed were further checked for distance and bearing from a point preceding 10 points upstream and the point succeeding 10 points downstream. If either distance or bearing change remained zero, the point was removed from the database.

Sparse Data: Some trips were found to have too few points for proper identification. The threshold ratio of distance to number of points was set such that speed between two consecutive points should not exceed 100 feet per second. If more than 50% of a trip consisted of points that did not match this criterion, the trip was discarded.

Noise Filtering: To filter points that are mainly signal jumps, a criterion similar to sparse data was used. If the distance from the point 10 steps before and/or 10 steps ahead

of the point being checked is such that it cannot be traversed in the time between the timestamps at a speed of 70 feet per second, then that point is removed from the dataset. An additional check, if a large group of 10 or more points are major deviations, was used to remove any GPS point that was over 5,280 feet from the point that is 10 positions prior to it.

Map Matching

There were two processes involved in the map matching part of the project. First, the network to which the GPS points to be snapped had to be cleaned and processed for the purpose of matching bicycling trips which are quite different from vehicular trips. For example, bicycling trips do not happen on freeways and keeping the freeways in the map might result in some nearby trips snapping wrongly to freeway segments. Therefore, we preprocessed the network map to better suit our purpose.

Three data sources were used to create the road network map. The Atlanta Regional Commission's street network shapefile (RC_ROUTES) was obtained from the travel demand modeling group of Atlanta Regional Commission (ARC). It is a modified version of the roadway database maintained by the Georgia Department of Transportation (GDOT) and focuses on state managed roadways rather than locally managed roadways and bikeways. However, it contains the most comprehensive inventory of roadway characteristics like speed limit, annual average daily traffic (AADT), number of lanes, truck volume etc. which are useful information for route choice modeling at a later stage. The second data source used is Open Street Map's (OSM) bicycle map for Atlanta. The OSM map has local roads and locally managed facilities which were not present in the RC_Routes map. The two maps were spatially joined based on a buffer distance to get a

more complete map of the road network of Atlanta. The resulting map was then cleaned for non-bicycling facilities like freeways. The final data source was the Metro Atlanta Bicycle Facility Inventory. The location of on street parking on roadways with conventional bicycle lanes and buffered bicycle lanes was manually coded in ArcGIS using Google Earth imagery. The treatment of intersection approaches with right turn only motor vehicle lanes that connect to links with conventional bicycle lanes, buffered bicycle lanes, or protected cycle tracks were also manually coded in ArcGIS using Google Earth Imagery. As a final measure, the trips were plotted on the map and checked for links traversed by cyclists but missing in the network. Such links were manually added where more than 2 bicycle trips were found to follow a path but the path was not marked as a link in the network. This was assumed to be mainly because of tendency and ability of bicyclists to use cut-thrus and private alleys which are not marked in regional network maps. However, shortcuts through parking lots were not added as links although there were multiple such cases.

Rebuilding the network file was done in ArcGIS partly because it was easier to merge multiple shapefiles in ArcGIS and also because we felt it was necessary to have a visual check on the merging and link imputation processes. The final shapefile was then imported to R using rgdal [shapefile to OGR] (Bivand et al. 2015) and shp2graph [OGR to igraph object] (Lu 2014) packages within R. The map was imported with the option of retaining all the associated properties as dataframes i.e. all the link characteristics were imported into R along with the shapefile. The graph object was then used for map matching using the packages igraph (Csardi et al. 2015) which is a fast and efficient package for handling large networks and spatstat (Baddaley et al. 2015) which is an

advanced spatial statistical package for analyzing spatial patterns. In the interest of computational time and tractability, after cleaning, the network was clipped to a 5-mile radius with the center at the intersection of Ponce de Leon and Monroe Avenue in downtown Atlanta. Figure 10 shows the study area under consideration.

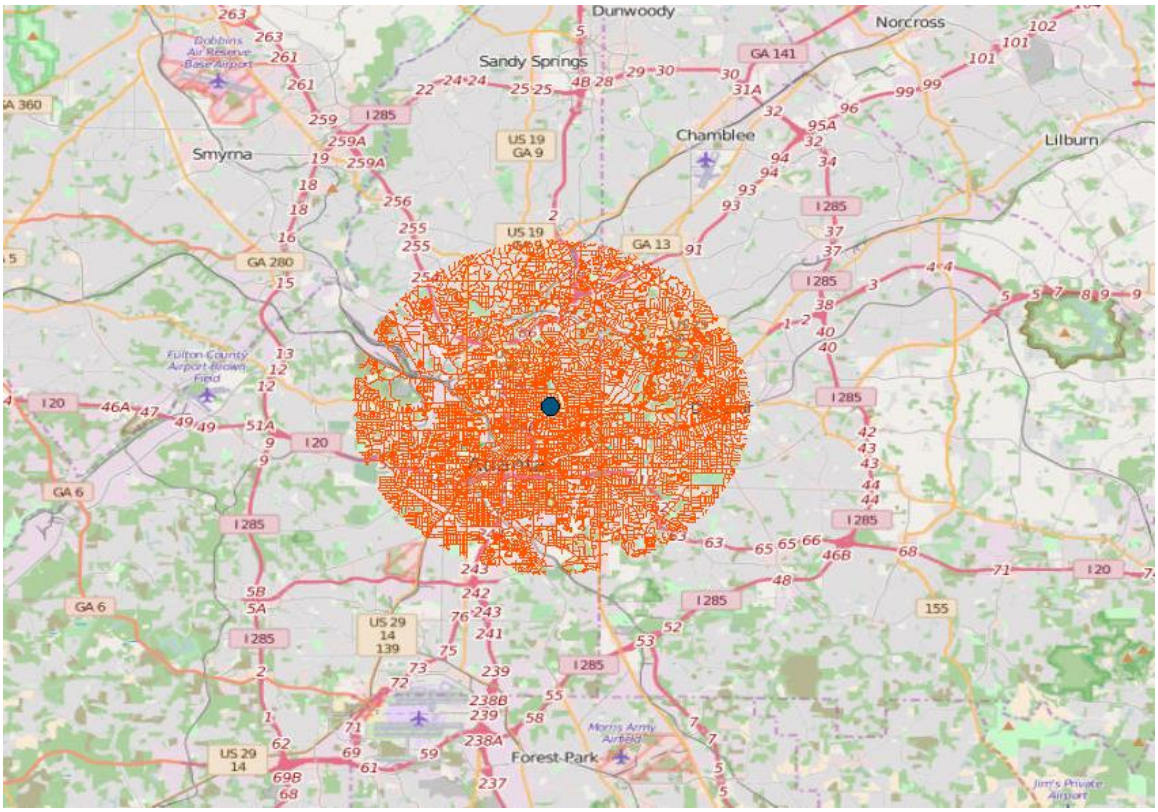


Figure 10. Study Area for the Dissertation

Once the network map was ready, we used two different methods to snap the trip data onto the network. The first base case used a combination of geometric and topological approach while the second method was designed to make use of the adjacency properties of network elements and reduce the computational burden of network search at every single instance of GPS recording.

Scenario 1: Base case

In this method, a spatial cross distance matrix between each trip point and each node in the network was created using '*spatstat*' package. The matrix was then sorted to get the minimum distance node for each trip point. The list of the nodes thus generated was first cleaned to keep one instance of a node occurring for multiple times *consecutively*. It should be noted that separate instances of the same node were retained in the list where sequences were broken by instances of other nodes occurring in between sequences of the same nodes. For example, a node sequence {4, 5, 5, 2, 5, 5} would be filtered via this process to {4, 5, 2, 5}. Then the list was filtered for oscillation of nodes – if a different node was visited after and before two instances of same node, that node was removed from the list i.e., after this step, the list showed earlier should become {4, 5, 5}. Finally, the list was filtered for unique values as one node could be the nearest node of multiple trip points and the list would contain multiple entries of that same node – at this stage, the list should only contain {4, 5}. The unique nodes were then *successively* checked for connectivity using the '*igraph'* package. As identification of a trip was dependent upon proper identification of the start point, a special check was introduced to identify the start point. The first node in the filtered list was checked to see if it was the start node for at least two consecutive links i.e. if at least three consecutive points after the first point were successively connected. Upon failure, the next point was checked for similar criteria and the process was repeated unless a match was found. In all cases, a match was found within the first three nodes. The connectivity check was then run successively on each node starting from the matched start node and the connected nodes were retained in a list. At any point during the process, a list was terminated when a point

was found to not be connected to any of its three consecutive nodes and a new list was started with the first point that was not connected to its preceding point. So, if the starting list was c = {1, 2, 5, 6, 7, 8, 9, 10}, node 1 was first checked if it was connected to node 2; if true, node 2 was then checked for connectivity with node 5 and then node 5 with node 6. If node 5 was not connected to node 6, it was checked if node 5 was connected to node 7 or node 8. If it was not connected to either node 6, node 7 or node 8, then list was terminated as c = {1, 2, 5} anda new list d = {6} was started. Next it was checked if node 6 was connected to node 7 – if true, then node 7 was added to list d such that d = {6, 7}and the check continued with addition of connected nodes to the list and termination of lists whenever one node had three consecutive nodes to which it was not connected. However, random checks on trips indicated that loss of connectivity either occurred at the beginning or at the end of the trips, implying that the points were GPS errors rather than actual trip points. The process would give multiple lists of connected nodes which could then be aggregated and used to find the actual trip once the last node was reached. Any list with 3 nodes or less was not considered for the purpose of aggregation as most of the road segments in the original map had more than 3 nodes (consisting of 2+ polylines). When aggregating, a check was introduced for a common node in the adjacency list of the last node of the first list and the first node of the second list and that node was added into the list between the other two nodes to get a continuous path. One of the biggest advantages of this method would be that it inherently adjusts for sparse data through aggregation of links and via imputation of the node from the adjacency matrix which could be missed due to sparse data. So, if there are two separate lists c = {1, 2, 5, 8} and d = {6, 7, 10, 11}, upon aggregation, the method checked the adjacency list of node 8 and

the adjacency list of node 6. If the adjacency of list of node 8 was {12, 13, 15} while the adjacency list of node 6 was {16, 13 and 9}, then node 13 was added to the end of list c and then list c and list d were aggregated.

The connecting links for successive pairs of nodes were then retrieved with their associated properties and stored as the chosen path for the trip.

Pseudo Code:

```
while trip-id ≠∅ {
d ←crossdist(points, nodes);
c←{for each col in d get row_id with min cell value};
filter c by:
  for i in c[n]
  {
   if {
c(i)=c(i+1)
 c[n]=c[n]-c(i+1);
       };
i=i+1;
   } end for
return c;

for i in c[m]
{
   if {
c(i-1)=c(i+1);
 c[m]=c[m]- c(i)
      };
i=i+1;
}end for
return  c;

filter c:
unique row_id;
return c;

check:
  for j in c[k]
  {
    if {
are.connected(c(j),c(j+1))==T
```

```
 j=j+1;
    elseif
       {
are.connected(c(j), c(j+2))==T,
check:
        if {
are.connected (c(j+1), c(j+2))==T
j=j+2;
            }
c[k]=c[k]-c(j+1);
j=j+1;
            }
elseif
{
are.connected(c(j), c(j+3))==T
check:
if{
are.connected(c(j+2),c(j+3))==T
   if{
(are.connected(c(j+2), c(j+1))==T
 j=j+3
     };
c=c[k]-c(j+1);
j=j+2
   }
c=c[k]-c(j+1)-c(j+2);j=j+1;
}
 j=j+3;
     }
counter=1
truncate c as list [counter]=c[j];
 counter=counter+1
}
 c =c[k]-c[j];
repeat process
}
Check:
 adjacency list of last element of list(i)
 with adjacency list of first element of list(i+1)
 p ←{common node};
final_list←unlist all lists
}
trip-id<- next(trip-id)
```

Validation:

Five random trips were chosen and the snapping algorithm was run on them. In all the five cases, the removed nodes were the last few nodes in the list. We also ran a Djikstra's shortest route between the first and last node before removal and in all the cases, the program could not find a valid route between them. However, when the same program was run to find the shortest route between the revised first and last node, it was able to find valid routes between them which indicates that removed points have a high probability of being GPS errors rather than valid trip identification points.

We did not have the issue of sparse data, so it was not possible to validate the effectiveness of the adjacency matrix approach for finding missing links. Since the proposed algorithm only checks for one missing node, it might have issues when there are multiple missing links in between two valid nodes. In such cases, it may be worthwhile to assume that the trip took the shortest path between the last node of the first list and the first node of the second list and then retrieved the nodes associated with that path. The final and complete node list then will consist of the two original lists with the nodes on the shortest path added in between the last node of the first list and the first node of the second list. Figure 11 shows examples of two trips matched to network links via this method. While the cleaning routine works very well and there is no random noise in the trip points, the matching is not yet 100% as the base network still consists of unconnected nodes and links.

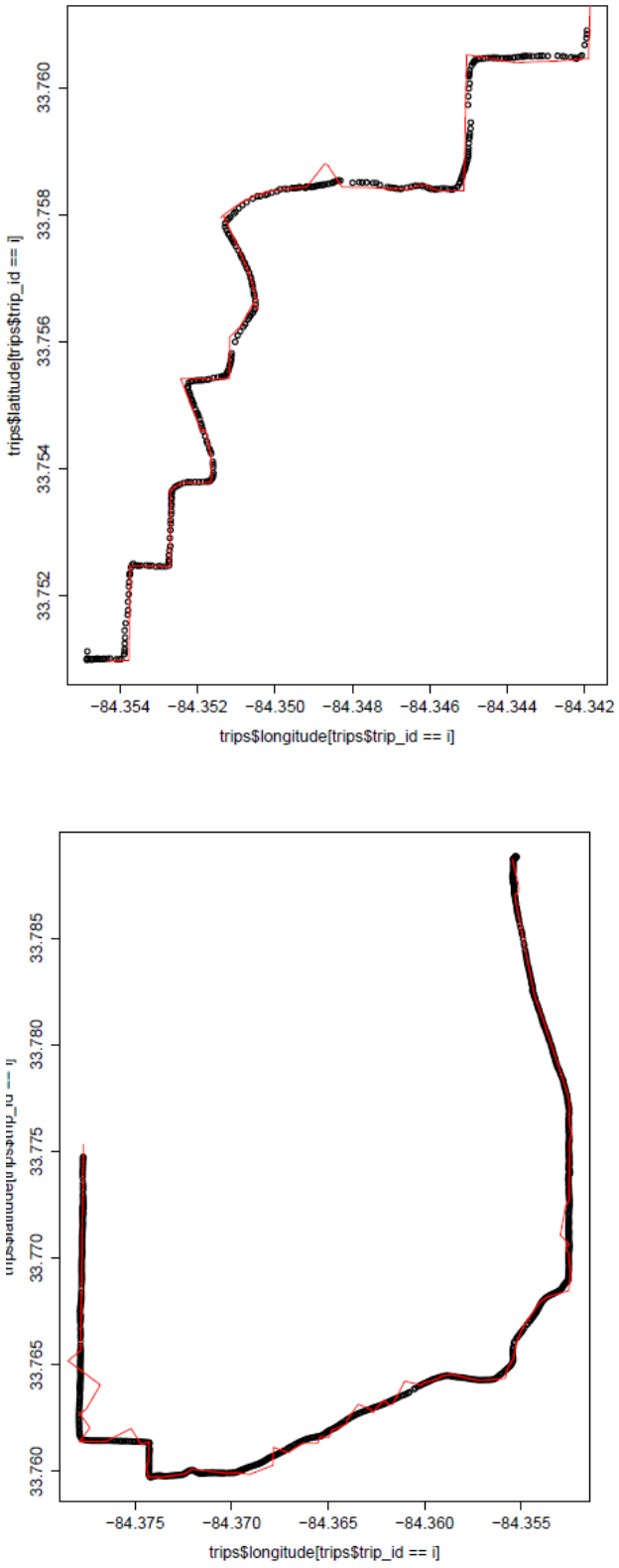


Figure 11. Example of Matched Network Links (in red lines) to GPS Trip Points (black transparent Circles)

Scenario 2: Map matching using adjacency matrix

In this case, first we created an adjacency matrix for all the nodes in the graph. Then, only the first five GPS points of a trip were selected and their nearest nodes were searched for in the entire network. The nearest nodes were then checked for connectivity and the first point that led to three consecutive connected links was flagged as the trip origin. We next created an adjacency list for the selected node and added the selected node to that list. For the next GPS point, we searched the nearest node from that list. Once that node was selected, we took the adjacency list of that node, added the node back itself and searched for the nearest neighbor for the next GPS point and the process was repeated until we reached the trip end. The final list was then first filtered for repeated occurrences of the same node consecutively and then it was filtered for instances where one different node occured between two instances of same node.

Pseudo Code:

```
while trip-id ≠∅ {
## get the start point ##
points_sub ←points[1:5]
d ←crossdist(points_sub, nodes);
c[m]←for each col in d get row_id with min cell value;
c[n]←filter c[m]by unique row_id;
check if c[n] is single element, then return that node
for i in c[n]{
a[j]←adjacency list(c(i));
check:
if c(i+1)is in a[j]== true;d[ ]=[c(i+1),..]
i=i+1 ;
elseif c(i+1) is in a[j]== false,
f[ ]=c(i+1)
 i=i+1;} end if
check:
if len [d]>len[f]
nodes=nodes-[f ]
else nodes=nodes-[d ]
## End Start Point ##
```

```
## Map Matching ##
for i in nodes {
a[j]←adjacency.list [nodes(i)];
c[k]←concatenate (a[j], nodes(i)]
check:
 min.distance(nodes(i+1), c[k]);
select c(k)withmindistance from nodes(i+1);
d[i]←{c(k),..}
i=i+1;} end for
 ## Post Processing ##
for i in d[i]{
 check:
if {
d(m)==d(m+1),
d[i]=d[i]-d(m);
m=m+1} endif
 return d};
for i in d[i]{
 check:
if {d(j-1)==d(j+1),
 d[i]=d[i]-d(j)
}
return d;}}
```

There are multiple advantages to the second method of map matching as compared to the first one. First, it is computationally efficient as it has to search the entire network only for the first few GPS points instead of all the GPS points in a trip. In the case of large network files like that of Atlanta, this is a huge advantage. Second, it does not require storing a huge distance matrix for further processing which renders the process memory efficient. Third, by using an adjacency list, it already asserts connectivity and no further check is required. Fourth, without any nearest distance threshold, it will always find an adjacent node for a GPS point which reduces the risk of prematurely ending a trip. The threshold distance value for the nearest node can be modified as needed for the research. Finally, the biggest advantage of this algorithm is that with slight modifications, it can be

used to probabilistically determine a route when GPS data are sparse. By adding link characteristics to the adjacency list, the algorithm can use the utility maximization or cost minimization concept to identify which link will be chosen by the user probabilistically. However, the problem with this method is that it is contingent upon correctly identifying the previous node. If any node identified within the process is wrong, there are chances that the trip will be identified completely wrong. This is particularly true if no threshold value for distance between the GPS point and the nearest node is set: in that case every GPS point will return a near node however far it may be and the trip will be a set of nodes that do not form a feasible path. On the other hand, this can also be seen as a second step of data cleaning where GPS points that are actually errors are forced to follow the actual route instead of having to do multiple checks as in the previous algorithm.

Conclusion

In this part of the research, a standardized data cleaning and map matching procedure was developed which can be useful for any related GPS based data collection effort. The code will be made open source and will be available for any future studies. It was found during this stage that having a complete and connected street network map is essential for proper execution of any snapping algorithm. Multiple platforms, both GIS based and script based, were tested and it was found that methods based on scripting languages like 'R' are computationally more efficient – however, there needs to be a procedure that will enable the results of scripts to be displayed in any GIS based software so that a visual check can be performed. In the future, this research will be undertaken by the current study's researchers.

CHAPTER 5

MODELLING CYCLISTS' CHOICE OF SHORTEST ROUTE AND DEVIATION FROM THE SHORTEST ROUTE

Introduction

Cycling trips are remarkably different from vehicular trips in their tendency to be on routes that maximize safety and comfort instead of routes that minimize travel time and/or distance (Dill and McNeil 2012, Buehler and Pucher 2012). The literature on cyclists and cycling trips suggest major safety perception factors to be high speed limits, high traffic volumes, last mile disconnect in the network, and an absence of physically separated facilities for cyclists (Dill and Carr 2003, Buehler and Pucher 2012). Earlier research by the authors on a stated preference dataset also showed that presence of separate infrastructure can positively influence the decision to adopt cycling across all age groups, particularly females and low confidence riders (Misra et al. 2014).

Building dedicated infrastructure, however, comes with substantial investment requirements and often with decisions to convert vehicular traffic lanes to bicycle facilities, both of which require strong justification that indicates potential benefits of such construction. It is, therefore, important for the planning agencies to know how far the cyclists are willing to travel to access a designated facility. Shafizadeh and Niemeier (1997), Aultman-Hall et al. (1998) and other researchers indicate that there is a threshold to the amount by which the cyclists are willing to deviate from their shortest distance-based route when accessing a facility that is off that shortest route. This observation is further complicated when we take into account the hypothesis that perceived safety varies

across cyclists depending often on their age, gender and/or experience (Dill 2004, Geller 2006).

One obvious way to solve this problem is to collect and analyze data on route choice behavior of existing cyclists. However, bicyclists being a small and distributed group, planning agencies often do not have sufficient funds to allocate for collecting data exclusively for bicyclists. Most studies related to infrastructure preference of cyclists are thus based on stated preference surveys where the users are asked to either rate or rank infrastructure attributes that may influence their decision to bicycle or to choose a particular route above another. One particular problem in such studies is segregation of theoretical preference from actual decision process. For example, while all cyclists rank traffic volume and traffic speed as having the most negative influence on the decision to cycle (Misra et al., forthcoming), it is not clear how many cyclists would avoid a short link with high traffic speed or volume if such a decision significantly increases their trip length or travel time. Studies where cyclists were asked to draw their route are vulnerable to recall bias and on participant's ability to remember – Aultman-Hall et al. (1998) had to discard ~ 20% of the trips because participants indicated impossible links and routes in the recall of their trips. In studies where participants are asked to choose from routes drawn on maps, they may often fail to get an accurate perception of impedances like slope or pavement condition and hence, may make a different choice when faced with an actual scenario rather than that shown in a stated preference survey. Further, since the choices are entirely created by the researchers it is difficult to ascertain if these alternatives would be actually considered by the participant in real life at all.

Revealed preference studies provide an alternative to the issues commonly faced within stated preference studies, but are data intensive. Recently, integration of GPS capabilities into hand held devices and smartphones have opened up a new dimension in low cost real time data collection, and bicycling research has gained significantly from such advances. In this chapter, revealed preference data collected via the smartphone application Cycle Atlanta (Misra et al. 2014) is used to understand the trip and route choice decisions of cyclists in Atlanta. The study is further motivated by the policy question of how far the cyclists are willing to deviate from the shortest network distance path to access facilities that they perceive to be safer than the shortest network distance path.

This chapter and subsequent two chapters are steps towards answering that question where we focus on understanding what factors influence cyclists in deciding a route choice and if those factors differ by cyclist age, gender or rider type. This particular chapter is a subset of that broader analysis and focuses on understanding whether a cyclist chose a shortest path or not and how far they are deviating from the shortest path to access perceived safer routes.

We first model the likelihood of choosing the shortest path between origin and destination as explained by presence or absence of certain road attributes like traffic volume, traffic speed, presence of bicycle facility etc., socio-demographic attributes of the cyclists and their confidence level. Next, we model the percent deviation from the shortest network distance based route as a function of difference in road attribute values between the network distance based shortest path and the chosen path. This part of the analysis is a step in the direction of understanding how far from the shortest distance the

cyclists will travel to access a facility – in its current form, this model gives an estimate of deviation over an entire path based on any individual factor or combination of factors. In the future, it may be worthwhile to extend this analysis to the link level which can give planners a better understanding of how far from an arterial a facility can be built and still be used by cyclists, without prior knowledge of the entire path that a cyclist might choose.

Two separate analyses are presented in this paper: in the first part, truncated and censored normal regression models are used and compared to model trip length as a function of socio-demographic attributes of the riders; in the second part, the cyclists' deviation from the shortest path is modelled using a two part sample selection model with the decision to not choose shortest path as the decision hurdle.

Literature Review

The literature review for this chapter is presented in two parts – the first part provides an overview of studies that looked into bicyclists' deviation from shortest network based distance, while the second part provides a brief background on truncated and censored regression models as well as on selection models – interested readers are referred to Cameron and Trivedi (2005), Greene (2005) and the classical text by Amemiya (1984) for an in depth and detailed coverage. It should be noted that most studies on deviation from shortest route presented here are stated preference studies and none of those studies made any distinction among cyclist types. It should also be noted that none of these studies actually developed models for cyclists' deviation from shortest route which is the focus of this chapter.

Cyclists' Choice of Shortest Route and Deviation from the Shortest Route

Multiple studies have looked into the route choice of cyclists and in almost all cases, cyclists were found to deviate from the shortest network distance path to access bicycle facilities or routes that they perceived to be safer (less traffic, more signalized, separate infrastructure) or more comfortable (less hilly, better pavement quality, better scenic route) (Shafizadeh and Niemeier 1997, Aultman-Hall 1998, Howard and Burns 2001, Hunt and Abraham 2007, Broach et al. 2010, Tilahun et al. 2007). A comprehensive literature review of factors influencing the decision to bicycle and bicyclist route choice can be found in the forthcoming article by Muhs et al. (2016) while some relevant studies on shortest route versus actual route are briefly covered in this section.

Shafizadeh and Niemeier (1997) conducted a stated preference survey among cyclists in Seattle at two locations and suggested that there exists a bikeshed of 0.4 – 1.2 miles around a separate bicycle facility and cyclists with trips within that bikeshed area are likely to deviate from the shortest path to access the facility. Aultman-Hall (1998) found that cyclists significantly deviated from the shortest route depending on grade, number of signals and turns, as well as frequency of buses on the shared route. A comparison between the shortest path and the actual route shows about 14 percent of the actual routes were the same distance as that of the shortest path while about 37 percent were within 0.1 km of the minimum distance, leaving almost half of the routes diverting a longer distance from the minimum distance. This suggests that cyclists have certain preferences in route choice that differ between their chosen path and the shortest distance path. Howard and Burns (2001) conducted a survey of 150 cyclists in Phoenix, Arizona

and concluded that actual route lengths are similar to shortest route by distance and time, although the actual routes tend to use more bicycle facilities. They inferred from the study that cyclists tend to choose routes that simultaneously optimize distance, time and safety but not any one objective singularly. Broach et al. (2010) found that 11 % of the commute trips and 12% of the non-commute trips are longer than their corresponding shortest network distances. Segadilha and Sanchez (2014) used GPS traces of bicyclists to understand their deviation from the shortest route and found that on an average, actual trips are 14.6% longer than the shortest route and about 83% of the cyclists chose paths that deviate no more than 20% by distance from the shortest network distance path. Segadilha and Sanchez (2014b) extended the study in the previous paper by asking the participants to use a GPS device for at least a week to collect the actual route traces that the cyclists used. Although the cyclists were frequently found to use routes that are longer than the shortest route, there was no significant difference between the route characteristics of the shortest route versus the chosen route. The authors mention that this may be due to the effect of averaging road characteristics over the entire route instead of segments.

Casello et al. (2010) designed a two part study for 100 cyclists in the Regional Municipality of Waterloo, ON, Canada. At the first stage the cyclists were asked to complete an online survey with demographic information and factors that influenced their propensity to bicycle. At the second stage, the cyclists were provided with a GPS device to record the routes they were using. A shortest route versus chosen route analysis was done using the same dataset. The results of the analysis reveal that travel time based cost functions or shortest distance measures are inadequate to capture the route choice

preferences of cyclists (Casello et al. 2010). A second study by Casello et al. (2011) extended the previous study to include 415 self-selected cyclists and showed that cyclists are forced to adopt a longer route where the land use pattern is not bicycle friendly. The study also concluded that the cost function (travel time and out of pocket cost) of cycling can be greatly reduced if trails are added to the existing infrastructure. A study on speed of cycling by El-Geneidy et al. (2008) showed that cyclists are able to travel at a slightly higher speed on off-street facilities than on on-street facilities. Since multiple studies have shown that travel time is an important factor in cycling decisions, ability to travel at a higher speed may translate to have a positive effect on a person's decision to bicycle although often partly counteracted by longer distances. Harvey et al. (2008) compared the shortest route versus the preferred route of the cyclists and showed that cyclists are willing to travel longer distances to feel safer, and as cyclists gain experience in riding along heavy traffic, they are less likely to choose a longer route than the shortest route. Krenn et al. (2014) studied the cyclists in the city of Graz, Austria which has a fairly high bicycle mode share (14%). The study consisted of data from two different studies – one conducted in 2005 where cyclists were asked to draw their most frequent bicycling route on a map and the second one, conducted in 2010, where the participants were asked to wear a GPS logger for 4 days so that their bicycling route could be recorded. Removing the participants who appeared in both the studies, data from 113 cyclists were used to understand the difference between the shortest route and the actual route adopted by the bicyclists. The sociodemographic variables considered in the study were age, gender, education level and BMI. Analysis showed that there is a statistically significant difference between the shortest possible route and the chosen route, although the

difference is not significant across age, gender, education or BMI. Cyclists chose bicycle paths and bicycle lanes over shortest distances along the main road without bicycle lanes. Additionally, chosen routes had significantly fewer traffic lights and crossings and are less hilly than the shortest route.

As mentioned previously, the majority of the studies reported in this section use stated preference surveys to understand the route preference of cyclists, a method which suffers from issues of recall bias and ideal choice scenario rather than actual choice. Additionally, none of the studies mentioned above modelled deviation from the shortest path as a function of path attributes or socio-demographic attributes of cyclists, although it is of paramount importance for the planning agencies to know how far the cyclists are willing to travel in order to access a facility. With limited infrastructure budget, planning agencies need substantial proof that their investment is not wasted on underused paths or paths that are beyond the acceptable range of deviation for a cyclist. This study addresses that information gap by modelling the deviation from shortest path as a two-step decision process.

Truncated Regression, Censored Regression and Selection Models

The econometric models used in this chapter are based on the concept that in behavioral studies, sampling methods are rarely completely random and there are almost always selection biases: sometimes intentional and sometimes unintentional. Intentional biases include studies that focus on a particular group rather than on a larger population; for example, studies that have data on disaggregate household income but decide to focus on households above a threshold income or studies that consider only people who smoke, people who voted for a particular party, etc. Unintentional bias can include examples

such as modelling number of hospital visits of a patient with insurance where data are obtained from the insurance provider – clearly, this is a subset of a larger population which is a mix of insured (both with the particular provider and with other providers) and uninsured patients, but the data in hand only reflects those registered with the insurance provider. *Data generated* via the first mechanism fall in the category of censored data where the analyst decides to group or exclude data below or above a particular level based on the requirements of the study, although there may be some disaggregate data available on the grouped variables. Censoring can happen both on the variable of interest and on the exogenous variables and is typically a many to one transformation process (Greene 2005). *Data obtained* via the second method fall in the group of truncated data where there is no information available to the analyst on the group complementary to the group under study or on the larger population of which the obtained dataset is a subset. Truncation typically happens only on the variable of interest or the response variable and is a *data collection issue* rather than *data generation issue* as in the case of censoring. Truncation also entails more information loss than censoring as censoring involves piling up all observations below or above the threshold *at the threshold value* while truncation involves *removing* all data at and below or above the truncation point. If truncation/censoring is such that observations below a certain threshold are lost, it is referred to as left truncation/censoring while if observations are lost above a certain threshold, it is referred to as right truncation/censoring. Both truncated and censored variable modelling assumes the presence of a latent continuous variable y^* which is observed as $y = y^*$ if $y^* > L$ for truncated (left truncated) and $y^* \geq L$ (left censored). For convenience, L, the truncation/threshold point is assumed to be zero but it can be

shown that any other value of L does not induce any additional inconsistency in model estimation as compared to the truncation/threshold point being zero (Greene 2005, Cameron and Trivedi 2005, Amemiya 1984).

Models for truncated and censored response variables can typically be classified into two methodological groups: (i) the two part model (2PM) and (ii) the sample selection model. The two part model, the first of the two groups, derives its name from the use of two different distributions for the two parts of the decision process: one to model the censoring/truncation mechanism and the second to model the observed outcome conditioned on it being observed. In other words, the two part model consists of two separate models, one which models whether the observation is below or above a threshold, or whether the observation belongs to a group or not and the second models the variable of interest for the observations which belong to that group above or below the specified threshold, depending on whether it is right or left censored. The second method for modelling truncated/censored data is called the sample selection method and uses a joint distribution for censoring mechanism and the response variable and then 'finds the implied distribution conditional on the outcome observed' (Cameron and Trivedi 2005, p 545). The sample selection model can be estimated either by a two step process where the censoring and the response variable are modeled using different distributions but linked through a control function (defined later in the text), or simultaneously by the method of maximum likelihood estimation.

The main aspect that requires special attention in modelling trip length data is that we do not observe trip length ≤ 0 as we do not observe the cyclists who did not make a trip. Therefore, the data is left truncated where we only observe those cases in which trip

length has a value > 0. Thus, we have complete data only on a subset of the population, i.e., we observe $y = y^* \; only \; if \; y^* > 0$ where y^* is the latent variable of interest. This is a sample selection problem with the further issue that we do not have information on what might have influenced someone's decision to participate in recording a trip. Therefore, we model the trip length as truncated normal distribution truncated at trip length = 0, jointly distributed with the latent propensity to make a trip which is only observed when trip length > 0.

On the other hand, in modelling deviation from the shortest path, it is important to recognize that there can be two distinct decision processes involved in the process, related or independent. A cyclist may first decide whether to choose the shortest network distance based path or not depending on trip length, traffic situation or even time of day. At the second stage, once the cyclist has decided to take a detour from the shortest path, the decision on how far to deviate may depend on link attributes like traffic speed or volume, pavement condition or presence of dedicated facility. We model the deviation from shortest path using both the two part model method as well as the sample selection method – the two part model gives the flexibility of modeling the deviation as two different decision processes where the first part is a binary probit model that gives the probability of whether a cyclist will choose the shortest path or not and the second part is a parametric logistic regression model that models the percentage deviation from the shortest path for the subset of cyclists who chose to deviate. The sample selection model, on the other hand, estimates the percentage deviation from the shortest path using a truncated regression model conditioned on the cyclist choosing to deviate from the shortest path. Although both the models use binary probit models to model the initial

decision, the major difference between the two methods is in modelling the deviation from the shortest path. While two part model uses a parametric logistic regression on a reduced dataset (estimated only on observations for which there is a deviation), the sample selection model uses a truncated regression model on the full dataset which, as shown earlier, gives a different estimate than parametric regression models. The two different methods of estimating sample selection models are presented in the following text.

To provide a general background of the estimation process of censored and truncated regression models, we start from the classical linear regression model where the latent variable y^* is linear in regressors with an additive error term that is homoskedastic and normally distributed, and present the functional form of the truncated normal regression followed later by the functional form of the two-part model and the sample selection model.

The standard linear regression model is given by:

$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ and

$$E[y^*|\boldsymbol{x}] = \boldsymbol{x}'\boldsymbol{\beta} \quad \textbf{(1)}$$

For truncated distributions, the observed y is defined as

$y = y^* \quad if\ y^* > 0$

$= missing \quad if\ y^* \leq 0$ (Cameron and Trivedi 2005)

The effect of ignoring any zero or negative value of the response variable pushes up the conditional mean of the response variable and reduces the variance. For example, for data left truncated at zero, assuming that the error term is not correlated with any explanatory variable, the conditional mean is given as

$$E[y|\boldsymbol{x}] = E[y^*|\boldsymbol{x}, y^* > 0]$$

$$= E[\boldsymbol{x}'\boldsymbol{\beta} + \varepsilon | \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0]$$

$$= E[\boldsymbol{x}'\boldsymbol{\beta} | \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0] + E[\varepsilon | \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon > 0]$$

$$= \boldsymbol{x}'\boldsymbol{\beta} + E[\varepsilon|\varepsilon > -\boldsymbol{x}'\boldsymbol{\beta}] = \boldsymbol{x}'\boldsymbol{\beta} + \sigma E[\varepsilon/\sigma|\varepsilon/\sigma > -\boldsymbol{x}'\boldsymbol{\beta}/\sigma]$$

$$= \boldsymbol{x}'\boldsymbol{\beta} + \sigma \frac{\phi\left(-\boldsymbol{x}'\boldsymbol{\beta}/\sigma\right)}{1-\Phi\left(-\boldsymbol{x}'\boldsymbol{\beta}/\sigma\right)}$$ where $\phi(.)$ and $\Phi(.)$ are the pdf and cdf of the standard normal distribution.

Denoting $\lambda = \frac{\phi\left(-\boldsymbol{x}'\boldsymbol{\beta}/\sigma\right)}{1-\Phi\left(-\boldsymbol{x}'\boldsymbol{\beta}/\sigma\right)} = \frac{\phi\left(\boldsymbol{x}'\boldsymbol{\beta}/\sigma\right)}{\Phi\left(\boldsymbol{x}'\boldsymbol{\beta}/\sigma\right)}$

in view of the symmetry about 0 of the standard normal distribution,

$$E[y|\boldsymbol{x}] = \boldsymbol{x}'\boldsymbol{\beta} + \sigma\lambda \quad (2)$$

where λ is called the inverse Mills ratio (Mills 1926, Cameron and Trivedi 2005).

Equation (2) shows that the mean of the truncated distribution is greater than the unconditional mean $\boldsymbol{x}'\boldsymbol{\beta}$ and that if the mean of a truncated distribution is calculated using linear regression, it will not include the term involving Mills ratio. Because that term is correlated with $\boldsymbol{x}'$, failing to include it (therefore leaving it in the error term of

the equation) will result in an endogeneity bias. Thus, the problems of fitting OLS regression equations to such truncated data are that (i) even though the underlying population mean is a linear function of x, the truncated mean will be non-linear and therefore, will need proper adjustment to be a consistent estimator of the population mean, (ii) linear approximation of the slope of the non-linear truncated mean via OLS regression estimation will be flatter and will lead to an inconsistent estimation of the slope parameter. Therefore, to estimate models where the variable of interest has a truncated distribution, a different analytical method other than linear regression is employed. The parameters are estimated using maximum likelihood where the log-likelihood is given by

$$\ln L = \sum_{i=1}^{N} ln\left[\frac{(1/\sigma)\phi\left((y_i - \boldsymbol{x'\beta})/\sigma\right)}{\Phi\left(\boldsymbol{x'\beta}/\sigma\right)}\right].$$

Amemiya (1984) named the truncated regression models estimated using the above mentioned log likelihood function as truncated Standard Tobit model, which is a variation of the Standard Tobit model developed for censored normal regression by Tobin (1953). It should also be noted that since this is nonlinear regression, the derivative of the conditional mean does not reproduce the parameter, but rather a factored version which is lower than the original parameter estimate (Greene 2005). For detailed derivation of first order and second order conditions of the log-likelihood maximization, readers are referred to Greene (2005), Amemiya (1984), Long (1997) and Cameron and Trivedi (2005).

In the second part of the analysis, a two-part model (2PM) and a sample selection model are used to understand how far the cyclists are willing to deviate from the shortest

network distance based path. Both the models consist of a discrete choice model (traditionally a probit model) at the first step which describes the boundary response choice and then, at the second stage, a parametric regression model is used to model the respondents that are not at the boundary response level. However, the sample selection model uses a truncated regression formulation for the second part while the two part model uses a standard regression formulation. As mentioned before, estimating a truncated dataset using standard regression gives a biased estimate as it does not include the term involving Mills ratio. However, for cases where the Mills ratio does not have a significant contribution to the model fit, the estimates from two part model and sample selection model are expected to be similar. Since sample selection models provide greater flexibility, two part models should only be used when the researcher is absolutely confident that there is no relation between the two processes being modelled and that the For the two part model, let the first stage model be, as Greene (2005) proposes

$$d^* = \boldsymbol{z}'\boldsymbol{\alpha} + u$$

$$T(d^*) = 1, (d^* > 0)$$

$$T(y^*|d^*) = y^* \; if \;\; d^* > 0,$$

$$= 0 \; otherwise.$$

At the second stage, the non-zero outcomes are modeled using the familiar formulation:

$$y^* = \boldsymbol{x}'\boldsymbol{\beta} + \varepsilon \qquad \text{where } \varepsilon \sim N(0, \sigma^2)$$

With this general regression model formulation, one major issue is that the predicted values of the response variable are not constrained to be within (0, 1). Since we model the deviation from the shortest path as a percentage of the shortest path, the response variable is bounded within (0,1). To address the issue, logistic regressions have been used in modelling the non-zero outcomes where

$E(y^*|x) = \frac{e^{-x\beta}}{1+ e^{-x\beta}}$, ensuring that $0 < E(y^*|x) < 1$.

For a bivariate sample selection model, let the latent variables of the selection and the outcome variables be (Cameron and Trivedi 2005)

$$y_1^* = x_1'\beta_1 + \varepsilon_1$$

$$y_2^* = x_2'\beta_2 + \varepsilon_2$$

Then, the observed selection and outcome variables are

$$y_1 = \begin{cases} 1 & if\ y_1^* > 0, \\ 0 & if\ y_1^* \leq 0 \end{cases}$$

$$y_2 = \begin{cases} y_2^* & if\ y_1^* > 0, \\ - & if\ y_1^* \leq 0 \end{cases}$$

The likelihood function of a bivariate sample selection model is given as

$$\mathcal{L} = \prod_{i=1}^{n} \{P[y_{1i}^* \leq 0]\}^{1-y_{1i}} \{f(y_{2i}|y_{1i}^* > 0) \times P[y_{1i}^* > 0]\}^{y_{1i}}$$

When ε_1 and ε_2 are uncorrelated, the two part model holds and the two processes can be modeled independent of each other. However, when they are correlated, the

maximum likelihood estimator assumes that the related errors are joint normally distributed and homoscedastic and estimates the likelihood function accordingly. The original Heckit estimator, given by Heckman (1976), on the other hand, involves estimating the model in two stages, and adds a term, obtained from the first-stage model, to the second-stage regression equation to control for the relation between the two stages:

$$y_{2i} = x'_{2i}\beta_2 + \sigma_{12}\lambda\left(x'_{1i}\widehat{\beta_1}\right) + \upsilon_i)$$

where $\widehat{\beta_1}$ is estimated from the first-stage probit model, λ is the estimated inverse Mills ratio mentioned earlier and υ is an error term. This formulation does not depend on the joint normality of errors assumption and hence is often preferred by practitioners over maximum likelihood estimation. We provide estimation by both of these sample selection approaches in the analysis section.

Data

Socio-demographic and Trip Related Variables

The primary data source for this research is the revealed preference trip data obtained from users of the GPS enabled smartphone application (app) Cycle Atlanta (Misra et al. 2014). The users of the app can voluntarily provide their socio-demographic, riding history, cycling frequency and trip purpose information, while the app, once switched on, collects the route trace of the bicycling trips they make at one GPS point per second. Once a trip is completed, upon receiving permission from the user, the GPS data are uploaded to a secure datacenter at Georgia Tech. Three separate databases are created from the data received from Cycle Atlanta users: (i) users - which stores user related information like socio-demographics, rider history and cycling frequency, uniquely tied

to user ids (ii) trips – which stores information on trip purpose, start and end GPS points, duration of the trip and any other input on the trips from the users, uniquely identified by trip id and user id and (iii) routes – which stores the complete GPS information for every trip including latitude-longitude of each point, timestamp, accuracy measures, speed and altitude, uniquely tied to trip id. Between October 2012 and June 2014, Cycle Atlanta had about 20,000 trips recorded by 1495 users, 60% of whom provided sociodemographic and cycling related information.

Network Data

Three data sources were used to create the road network map. The Atlanta Regional Commission's street network shapefile (RC_ROUTES) was obtained from the travel demand modeling group of Atlanta Regional Commission (ARC). It is a modified version of the roadway database maintained by the Georgia Department of Transportation (GDOT) and focuses on state managed roadways rather than locally managed roadways and bikeways. However, it contains the most comprehensive inventory of roadway characteristics like speed limit, average annual daily traffic (AADT), number of lanes, truck volume, etc. which are useful information for route choice modeling at a later stage. The second data source used was Open Street Map's (OSM) bicycle map for Atlanta. The OSM map has local roads and locally managed facilities which were not present in the RC_Routes map. The two maps were spatially joined based on a buffer distance to get a more complete map of the road network of Atlanta. The resulting map was then cleaned for non-bicycling facilities like freeways. The final data source was the Metro Atlanta Bicycle Facility Inventory. The location of on street parking on roadways with conventional bicycle lanes and buffered bicycle lanes was manually coded in ArcGIS

using Google Earth imagery. The treatment of intersection approaches with right turn only motor vehicle lanes that connect to links with conventional bicycle lanes, buffered bicycle lanes, or protected cycle tracks were also manually coded in ArcGIS using Google Earth Imagery. As a final measure, the trips were plotted on the map and checked for links traversed by cyclists but missing in the network. Such links were manually added where more than 2 bicycle trips were found to follow a path, but the path was not marked as a link in the network. This was assumed to be mainly because of tendency and ability of bicyclists to use cut-thrus and private alleys which are not marked in regional network maps. However, shortcuts through parking lots which were used to skip intersections and not links, were not added as links as we did not have information on the intersections to understand why cyclists were avoiding them.

Variable Definition

Two different sets of variables were used in these models – infrastructure related variables and sociodemographic variables. Accordingly, in this section, we present a description of all the variables used in two subsections.

Infrastructure Related Variables

All infrastructure related variables used were in the form of link level attributes, attached as a data table to the street network map obtained from ARC. Since this research is not modelling route choice decision as being an aggregate of multiple link level decisions, a weighted average approach was used to convert link level variables to route level variables. Table 9 shows the link attribute variables, their weighting factors and their route level aggregation methods that were used for this analysis. The principal idea behind using weighted average of variables was to understand the effect of each

attribute as a function of the link length. Therefore, all attributes were weighted by the link length and then averaged at route level by the route length. Brief definitions of the variables used are provided below:

Annual Average Daily Traffic (AADT) is a standard average measure for volume of traffic on a link in a street network. It is calculated as

$$AADT = (Total\ yearly\ traffic\ volume\ on\ the\ link)/365$$

For route level analysis in this paper, route AADT is calculated as AADT on a link multiplied by the ratio of link length to total length of the path and then summed over all links that constitute the path (see Table 9).

Speed limit is the posted speed limit for each link and not the actual speed of the vehicles travelling on that link. This is also a link level variable that is converted to a route level variable by weighting it by ratio of link length to path length and summing over all links that are part of the path (Table 9).

Number of lanes in its given form was the number of lanes on the link in one direction. To convert this to two-directional for this analysis, the variable was made into a ranked categorical variable with 3 levels: 1, if the number of lanes on the link ≤ 2; 2, if number of lanes = 3 or 4 and 3 if the number of lanes > 4. The logic behind this conversion was that since we did not have direction attached to the lane data, for a 2 lane link, the worst case for a bicyclist is 1 lane in each direction while for 3 or 4 lanes, the worst case is 2 lanes in any one direction or both directions. Similarly, for roads with more than 4 lanes, the bicyclists may have to face 3 lanes in any one direction. Roads with more than 3 lanes in one direction were not expected to be commonly used by

cyclists. The variable is converted into a route level variable by the same method as used for AADT and speed limit (Table 9). It should be noted that turn lanes were not marked in the network data file, so it was not considered for this analysis, although it should be taken into account in future analysis.

Slope was calculated as the elevation difference between the start node and end node of each link. Elevation of the nodes were obtained via API calls to Google maps wherein latitude and longitude of each node of the basemap were used as input. The link level difference in elevation was multiplied by the ratio of link length/path length and summed over all links to convert it into a route level variable (Table 9). A caveat to averaging slope over length of route is that it is possibly not a realistic representation of how slope affects a cyclist's decision to choose a route. A significantly high slope even for a short segment might deter a cyclist from choosing that link while when that slope is averaged over the route length, it might not be that significant. In future, some other methods need to be devised for an actual representation of the effect of slope on cyclists.

Presence of Bicycle Facility was coded in as binary variable for each link. Specifically,

$\boldsymbol{\delta_b}$ **= 1** if the link had a bicycle facility and

$\boldsymbol{\delta_b}$ **= 0** if the link did not have a facility**.** No distinction was made between on or off road facilities or among different types of facilities as any facility other than bike lanes would have been limited enough to form too small a sample for analysis. To use the variable at route level, the binary variable was multiplied by link length and then summed over all links and averaged over path length (Table 1). The resulting route level variable was the

fraction of the path that has bicycle facilities, which is an important determinant of route choice for cyclists.

Presence of Sidewalks was also coded as a binary variable indicating whether the link has a sidewalk or not, irrespective of whether it is on the left side, right side, or both. This variable was included in the analysis to understand if having a sidewalk helped a cyclist feel safer, even though riding on the sidewalk is not legally allowed for cyclists in the study area. The link level variable was transformed into route level variable using the same method as with the presence of a bicycle facility.

Traffic Stress was a variable created based on our previous research which showed that traffic speed and volume are correlated as factors that affect any bicycling decision (Misra et al., forthcoming). It was postulated that cyclists are less averse to links that have one of the two factors but are severely deterred by the combined effect of both. That is, whereas cyclists might be inclined to choose links with heavy traffic but low speed or links with high speed but low traffic if that link significantly reduces travel distance, they are most likely to avoid any link that has both heavy traffic and a high traffic speed, even if that link is on the shortest path. Therefore, a link level interaction term was created as AADT× speed limit and was named traffic stress. The variable was multiplied by the ratio of link length to path length and summed over all links of the path to convert it to a route level variable.

Table 9. Infrastructure Related Variables Used in Models and their Modifications

Attributes for link *a* of path *k*	Base Value	Link Level Weighted Value	Route Level Aggregated Value
Link Length	l_a	l_a	$\sum_{a \in \Gamma_k} l_a$
Annual Average Daily Traffic (AADT)	$(aadt)_a$	$(aadt)_a \times \frac{l_a}{L_k}$	$\sum_{a \in \Gamma_k} (aadt)_a \times \frac{l_a}{L_k}$
Speed Limit	s_a	$s_a \times \frac{l_a}{L_k}$	$\sum_{a \in \Gamma_k} s_a \times \frac{l_a}{L_k}$
Number of Lanes	n_a ($ranked\ categorical$)	$n_a \times \frac{l_a}{L_k}$	$\sum_{a \in \Gamma_k} n_a \times \frac{l_a}{L_k}$
Slope	$m_a = m_j - m_i$ ($m_i = elevation\ at\ start\ node$, $m_j = elevation\ at\ end\ node$)	$m_a \times \frac{l_a}{L_k}$	$\sum_{a \in \Gamma_k} m_a \times \frac{l_a}{L_k}$
Percent of Truck	p_a	$p_a \times \frac{l_a}{L_k}$	$\sum_{a \in \Gamma_k} p_a \times \frac{l_a}{L_k}$
Presence of Bicycle Facility	δ_b $Binary(0,1)$	$\delta_b \times l_a$	$\frac{1}{L_k} \sum_{a \in \Gamma_k} \delta_b \times l_a$
Presence of Sidewalk	δ_{sw} $Binary(0,1)$	$\delta_{sw} \times l_a$	$\frac{1}{L_k} \sum_{a \in \Gamma_k} \delta_{sw} \times l_a$
Traffic Stress	$ts_a = (aadt)_a \times s_a$	$ts_a \times \frac{l_a}{L_k}$	$\sum_{a \in \Gamma_k} ts_a \times \frac{l_a}{L_k}$
Path Overlap Correction (Bovy 2008)	PSC_k	$\frac{l_a}{L_k} ln \sum_{l \in C} \delta_{al}$	$\sum_{a \in \Gamma_k} \frac{l_a}{L_k} ln \sum_{l \in C} \delta_{al}$

Socio-demographic Variables

The other set of variables used in this analysis are the socio-demographic attributes of the cyclists. A detail of discussion of all the socio-demographic variables is presented in Chapter 3 and readers may refer to that chapter for related information. The variables under consideration were age, gender, rider type, cycling frequency and rider experience. The route choice study on cyclists in San Francisco (Hood et al. 2011) found that cycling frequency affected route choice while Dill and McNeil (2012) found rider experience affected comfort and confidence with different street infrastructures which

may eventually influence their decision to choose a route. From previous research (Misra et al., forthcoming) cycling frequency, rider experience and gender were found to be significant determinant of cyclist confidence, measured by their self-classification into different rider types. This significantly simplifies the model building step as well as makes the model parsimonious where including only rider type accounts for effects of both cycling frequency and rider experience. Additionally, these are variables that require separate data collection efforts than what is easily and publicly available. Developing a model with such variables as explanatory variables will imply requiring data on those variables for any future use, which may not be desirable. Therefore, the ability to replace two such variables with one provides greater flexibility for future use of the models.

For every model reported, three models were experimented with in the following sequence – one with only rider type as the individual specific variable, one with age and rider type as individual specific variables and the last one with age, rider type and gender as individual specific variables. In previous research, gender was also found to be correlated to rider type but was included in one of the models along with rider type to identify any gender specific idiosyncrasy that was not identifiable via rider type only.

Multivariate Analysis

As of June 2014, 18,657 trips had been recorded via the Cycle Atlanta app by 1495 users. After clipping the trips to be within a 5-mile radius of a chosen intersection point (Figure 7), the number of trips was reduced to 15,463. Of these 15,463 trips, 9,537 trips were commute trips (62%), 577 trips were school trips (4%), 2,461 trips were exercise trips (16%) and the remaining were errands, shopping and 'other' trips (18%) (Figure 12 (b)). Figure 12(a) shows the number of trips recorded by each rider. The

median number of trips recorded by one user is 4 while the mean number of trips recorded by a single user is 12, indicating a heavily right skewed distribution. A possible explanation is that a handful users of Cycle Atlanta record a significantly higher number of trips than an average user which is evident from users recording trips from anywhere between 1 and 500 in number. Figure 12(f) shows the frequency distribution of trip length. The mean commute trip length, marked by a dashed red line was found to be about 3.9 miles (about 5.5 kms) while the median commute trip length is 3.6 miles.

The trip data were then linked to the sociodemographic data via the user id key. Of the 1495 users, only 989 users provided information on rider type and 932 users had complete information on age, gender and rider type which were the three socio-demographic variables to be considered for further analysis. Therefore, only trips made by those 932 users were retained in the dataset after the initial analysis which reduced the number of trips from 15,463 to 8,323 and the number of commute trips from 10,114 to 5,222. Figure 12(c) shows the trip purpose across age – riders in the age group > 45 years use cycling for exercise more than any other group. Figure 12(d) shows trip purpose by gender, and since the data are heavily dominated by male cyclists, they are the dominating group in all trip purpose categories. However, female riders have almost a similar share of shopping trips in spite of being a small fraction of the riders. This calls for particular consideration in land use planning to allow women to do trip chaining comfortably and easily. Figure 12(e) shows trip purpose by rider type. The strong and fearless riders make more social and shopping trips by cycling than other types of riders, while enthused and confident riders use cycling for commute more than any other rider type. Comfortable but cautious riders use cycling for exercise more than other rider types.

Figure 12(g)(i) and Figure 12(g)(ii) show trip length by age. Figure 12(g)(i) shows the distribution of age across different trip lengths considering all trip purposes. Consistent with Figure 12(c), riders in the age group 45+ are more represented in trip lengths greater than 15 miles that are actually exercise trips and are not seen in Figure 12(g)(ii) which shows the same distribution but only for commute trips. It should be noted though that the highest frequency of trips for younger riders are at a shorter distance than that of senior riders which is initially counterintuitive. However, one of the reasons may be that senior riders are less likely to choose shorter routes if that does not provide sufficient safety and comfort while younger riders may prefer shorter distance to a detour for a bike facility. The younger riders are also dominated by college students, and their commutes may be much shorter in length. Figure 12(h) shows trip length across gender, and we see a similar trend as age here – the highest frequency of trip lengths for women are longer than that of men. Figure 12(i) shows trip length across rider type, and enthused and confident riders are seen to have shorter trips than comfortable but cautious riders. Strong and fearless riders have slightly longer mean trip length than enthused and confident riders, but that may be because they bicycle longer distances.

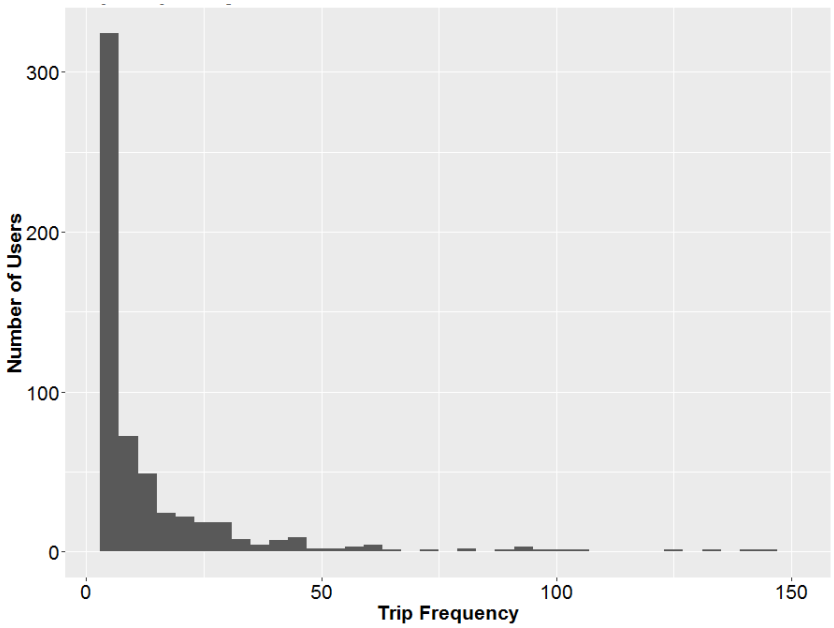


Figure 12(a) Number of Users by Trip Frequency

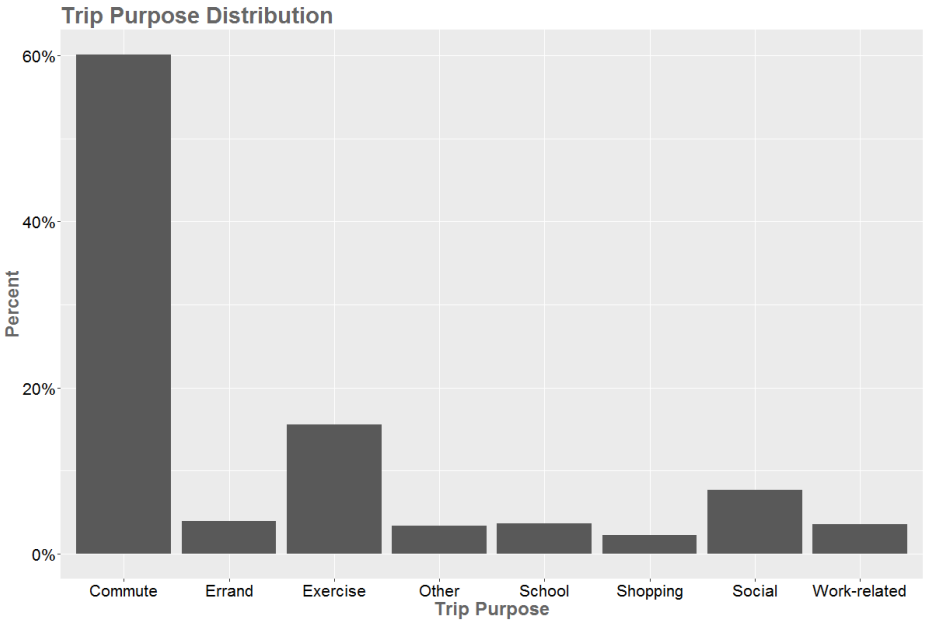


Figure 12(b) Trip Purpose Distribution

Figure 12(a)-(i). Cycle Atlanta Trips: (a) Number of Trips Recorded by Users (b) Trip Distribution by Purpose (c) Trip Purpose Distribution across Age (d) Trip Purpose Distribution across Gender (e) Trip Purpose Distribution across Rider Type (f) Trip length Distribution (g) Trip Length across Age (h) Trip Length across Gender (i) Trip Length across Rider Type

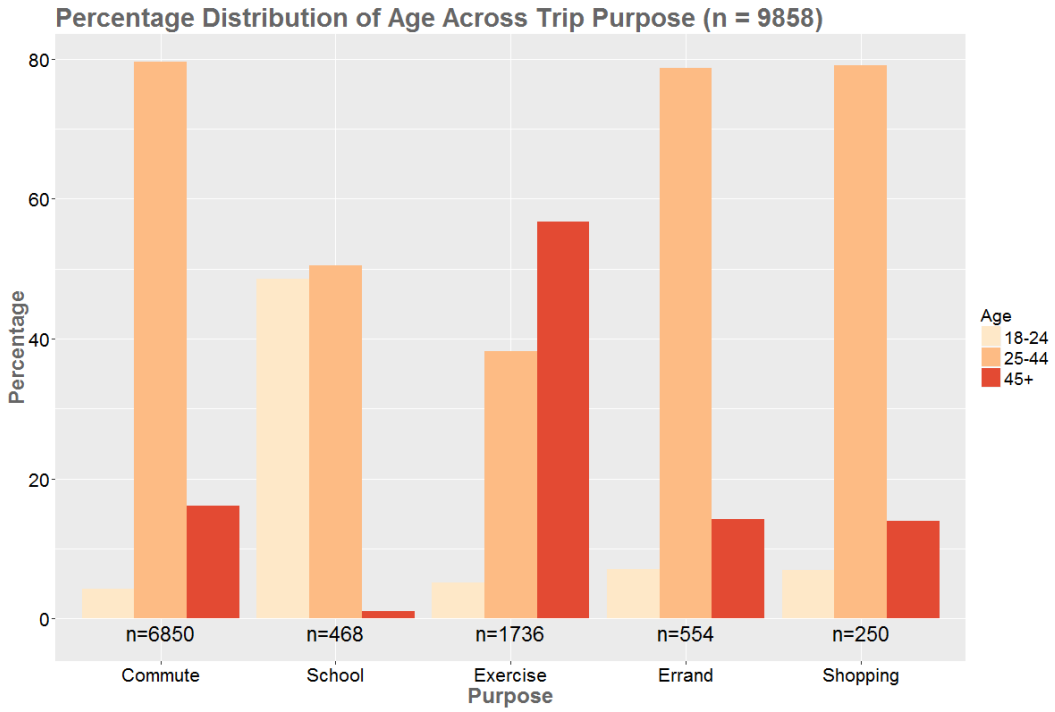


Figure 12(c) Trip Purpose Distribution across Age

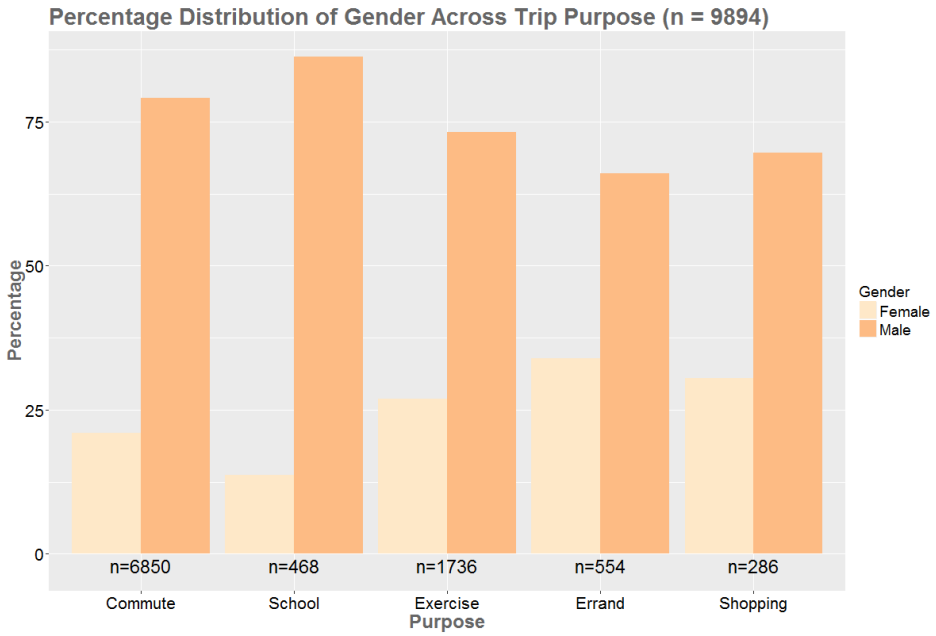


Figure 12(d) Trip Purpose Distribution across Gender

Figure 12(a)-(i). Cycle Atlanta Trips Continued

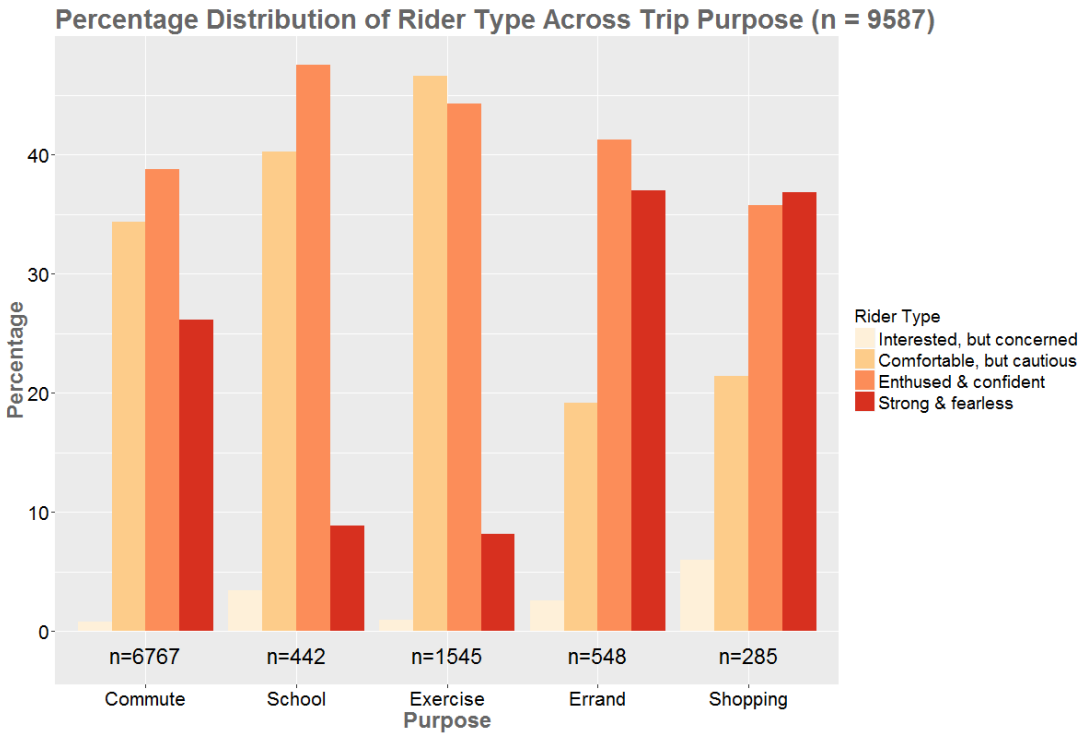


Figure 12(e) Trip Purpose Distribution across Rider Type

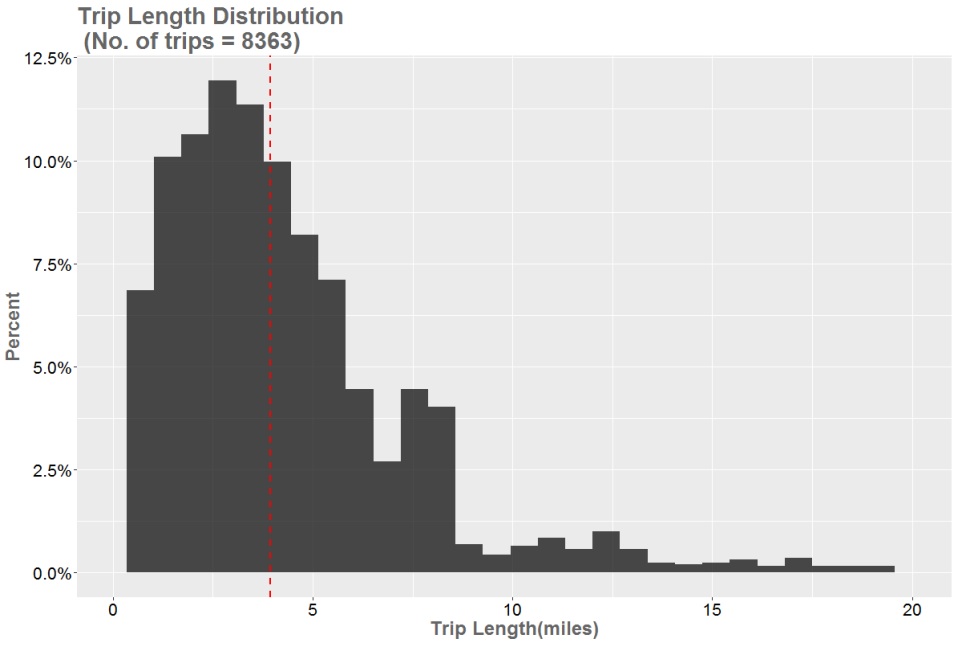


Figure 12(f) Trip Length Distribution

Figure 12(a)-(i). Cycle Atlanta Trips Continued

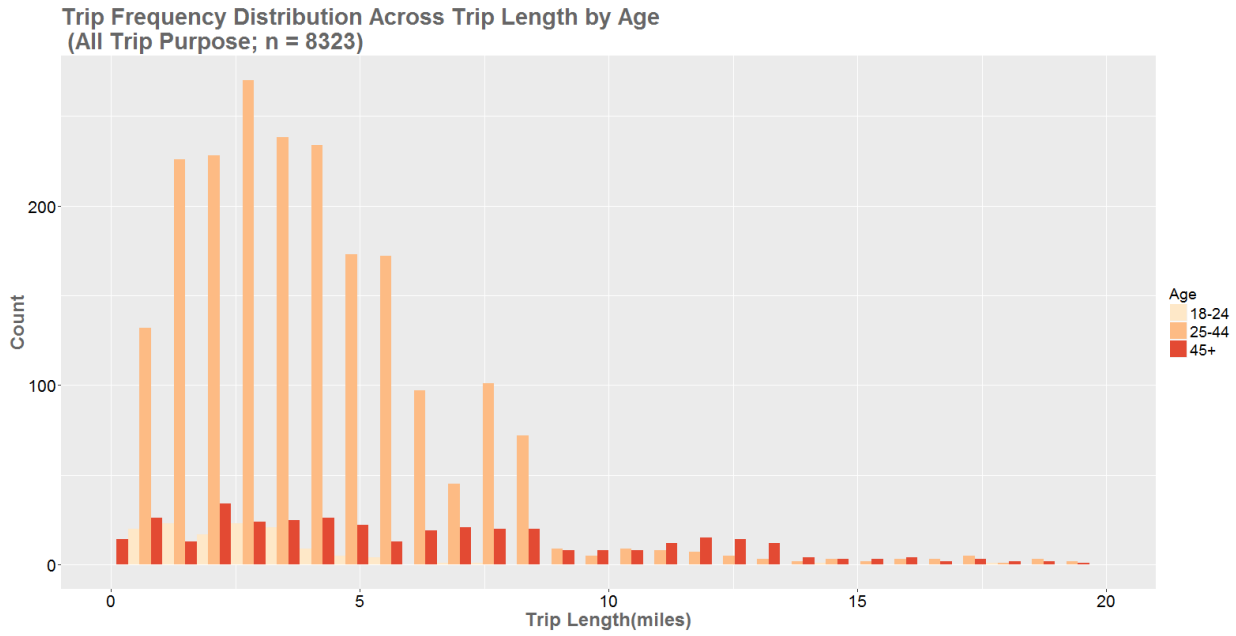


Figure 12(g)(i)(a) Trip Frequency across Trip Length by Age Groups for All Trip Purpose

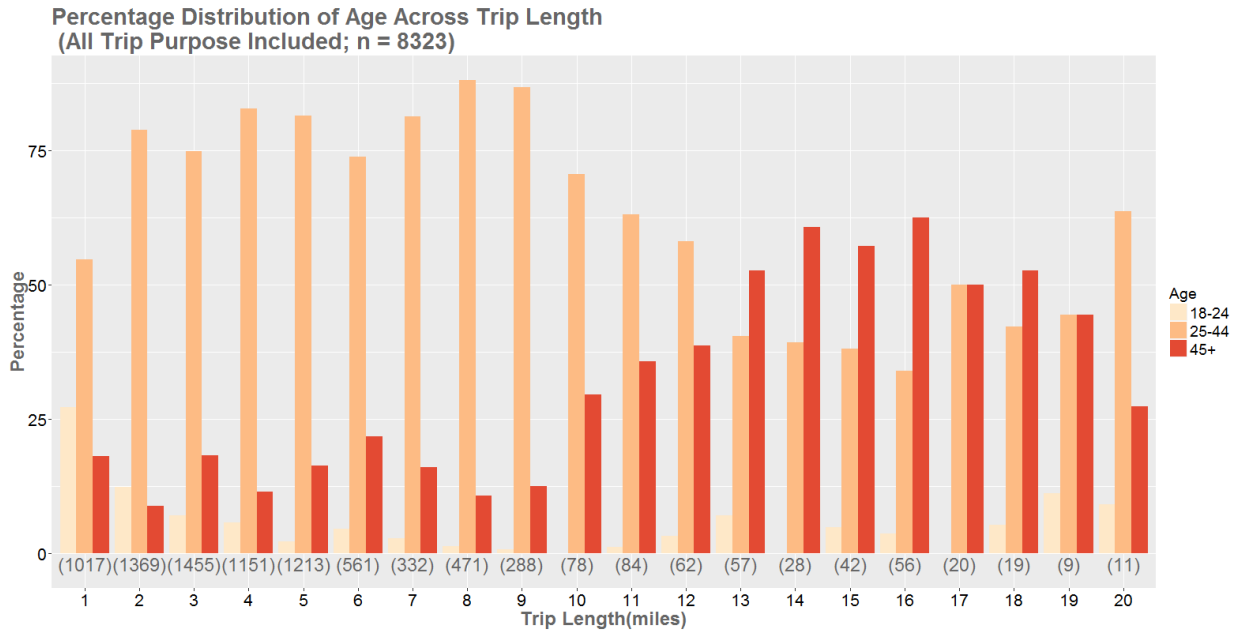


Figure 12(g)(i)(b) Age Group Distribution across Trip Length for All Trip Purpose

Figure 12(a)-(i). Cycle Atlanta Trips Continued

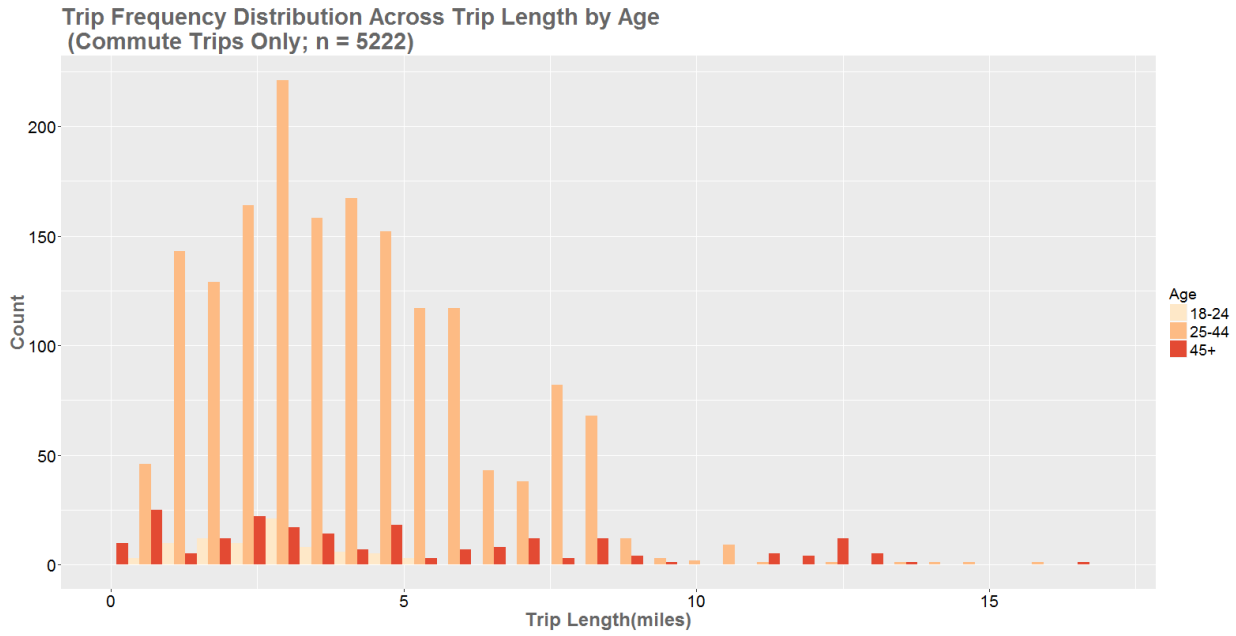


Figure 12(g)(ii)(a) Trip Frequency across Trip Length by Age Groups for Commute Trips

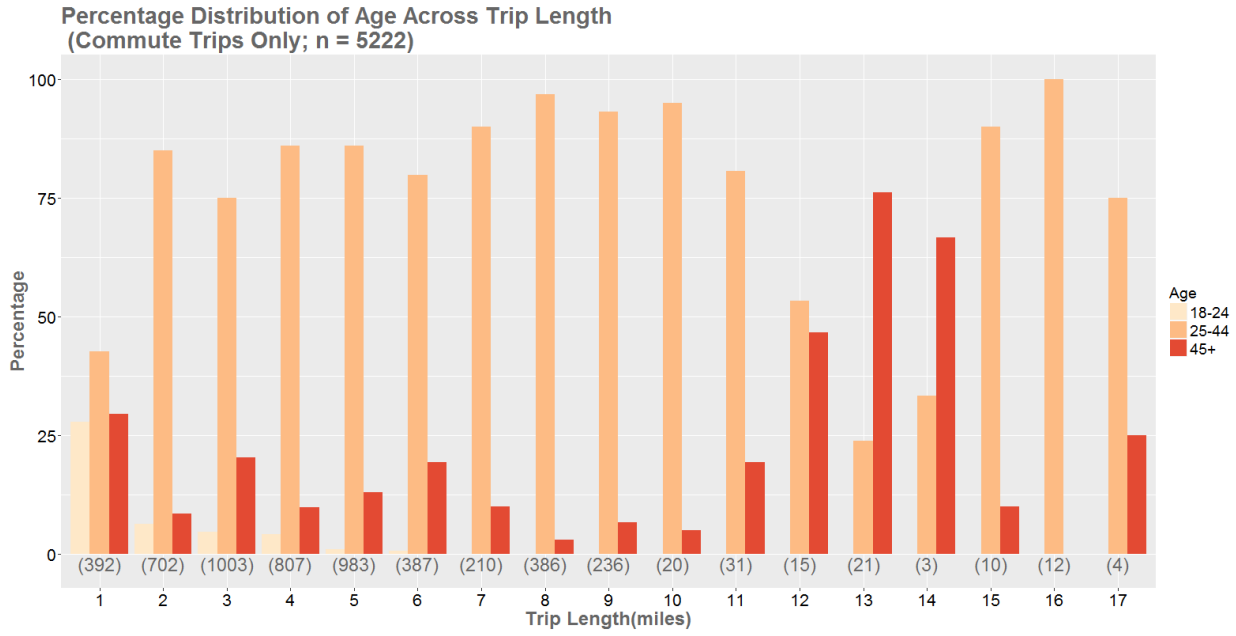


Figure 12(g)(ii)(b) Age Group Distribution across Trip Length for Commute Trips

Figure 12(a)-(i). Cycle Atlanta Trips Continued

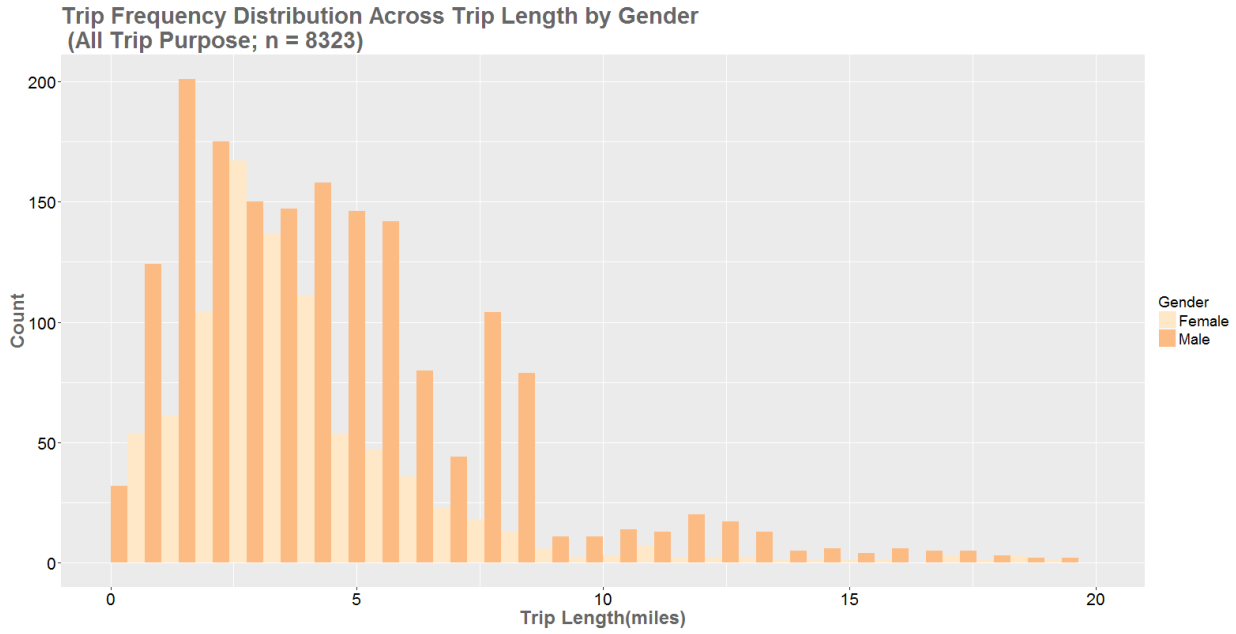


Figure 12(h)(i)(a) Trip Frequency across Trip Length by Gender for All Trip Purpose

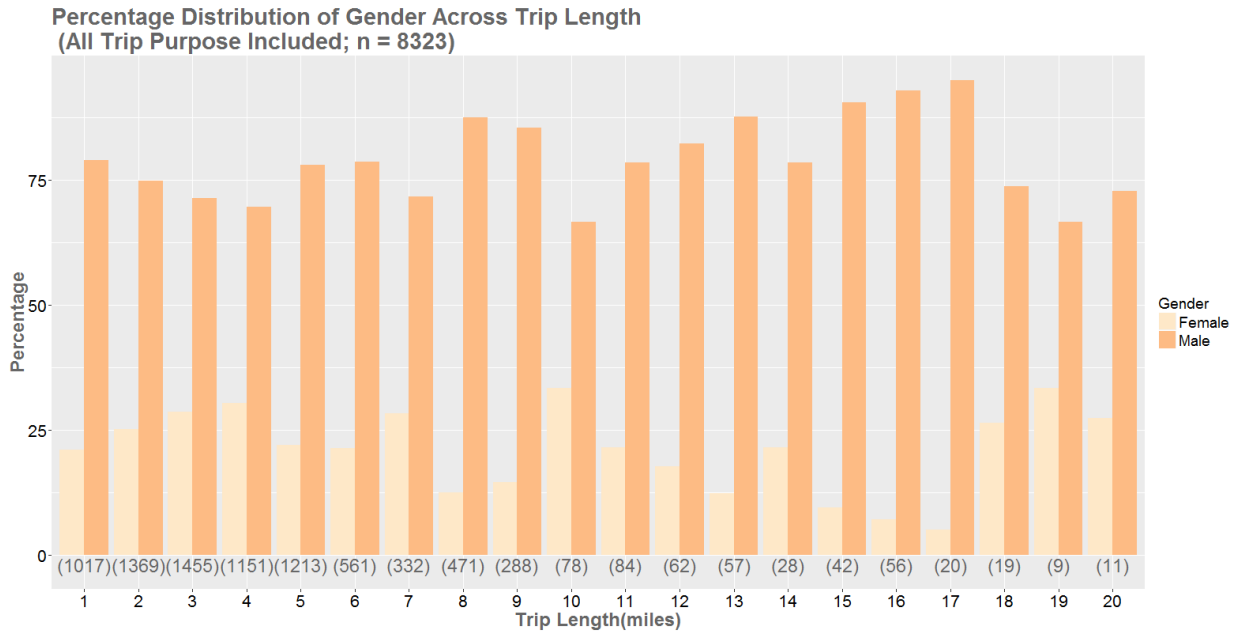


Figure 12(h)(i)(b) Gender Distribution across Trip Length for All Trip Purpose

Figure 12(a)-(i). Cycle Atlanta Trips Continued

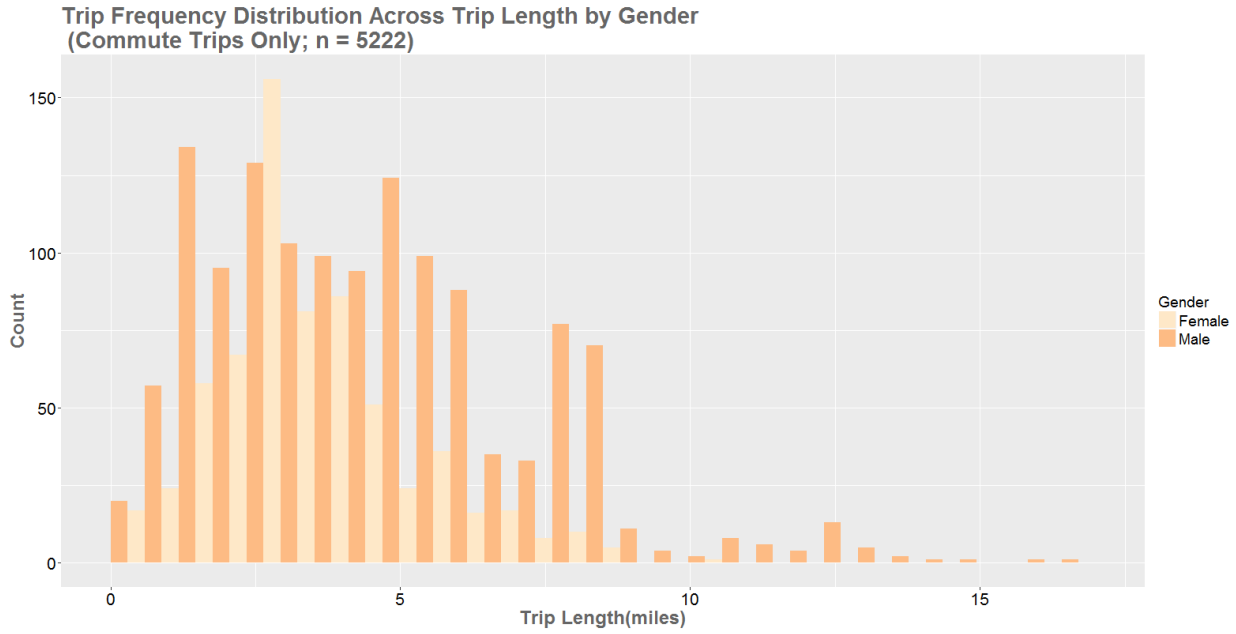


Figure 12(h)(ii)(a) Trip Frequency across Trip Length by Gender for Commute Trips

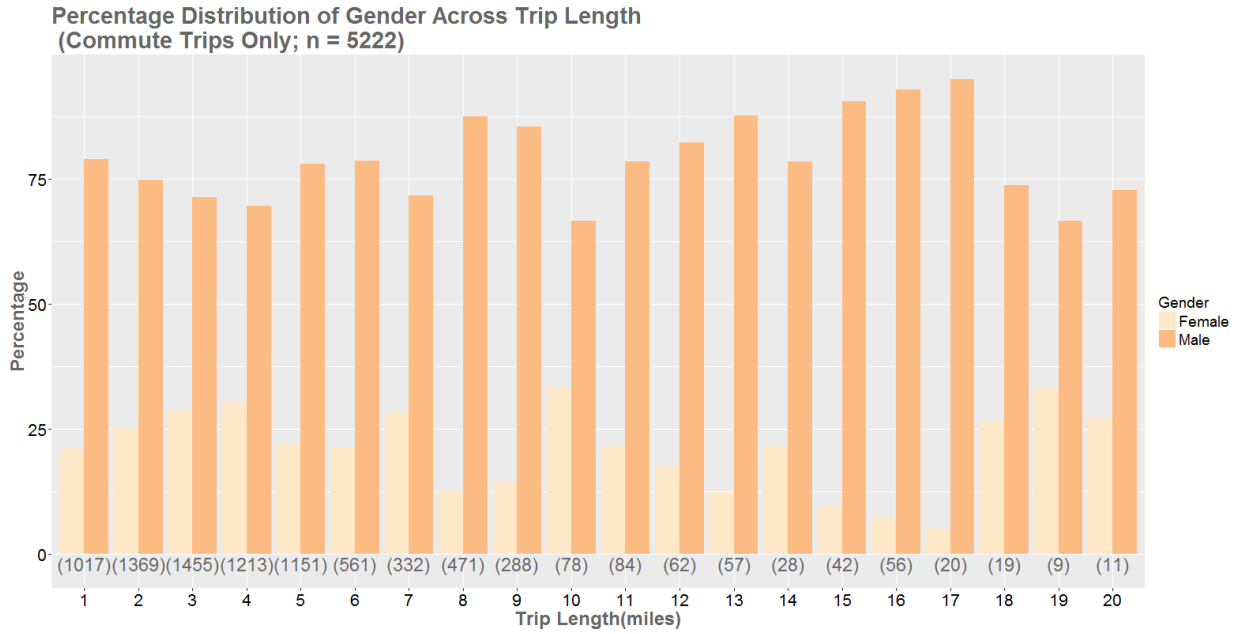


Figure 12(h)(ii)(b) Gender Distribution across Trip Length for Commute Trips

Figure 12(a)-(i). Cycle Atlanta Trips Continued

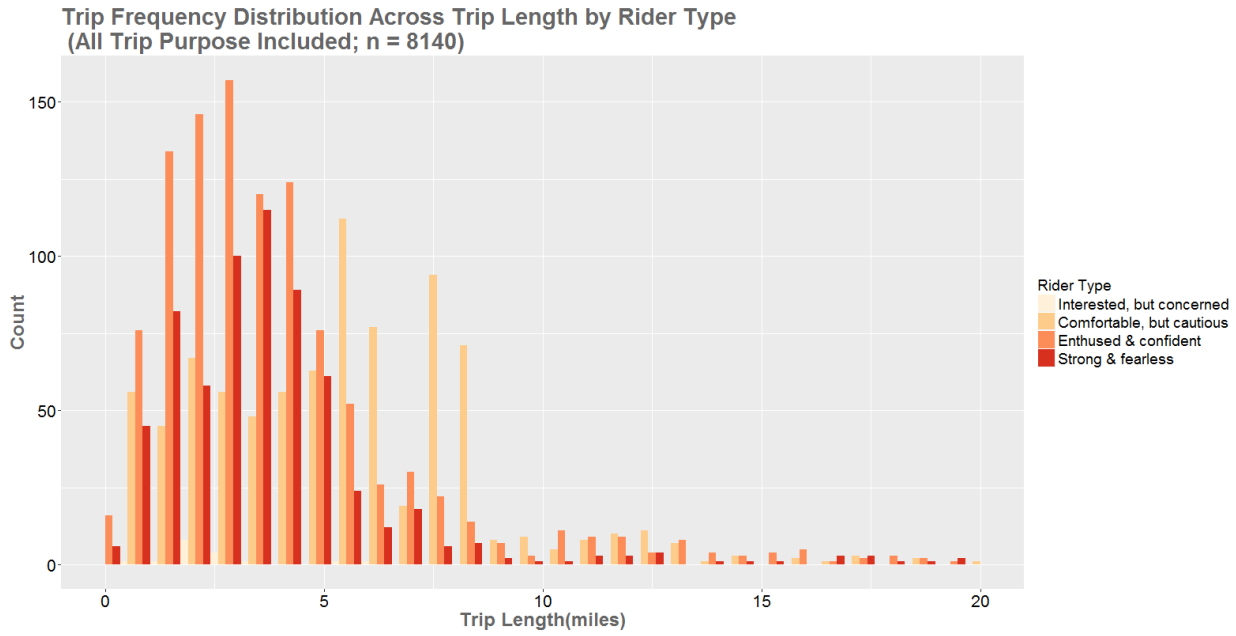


Figure 12(i)(i)(a) Trip Frequency across Trip Length by Rider Type for All Trip Purpose

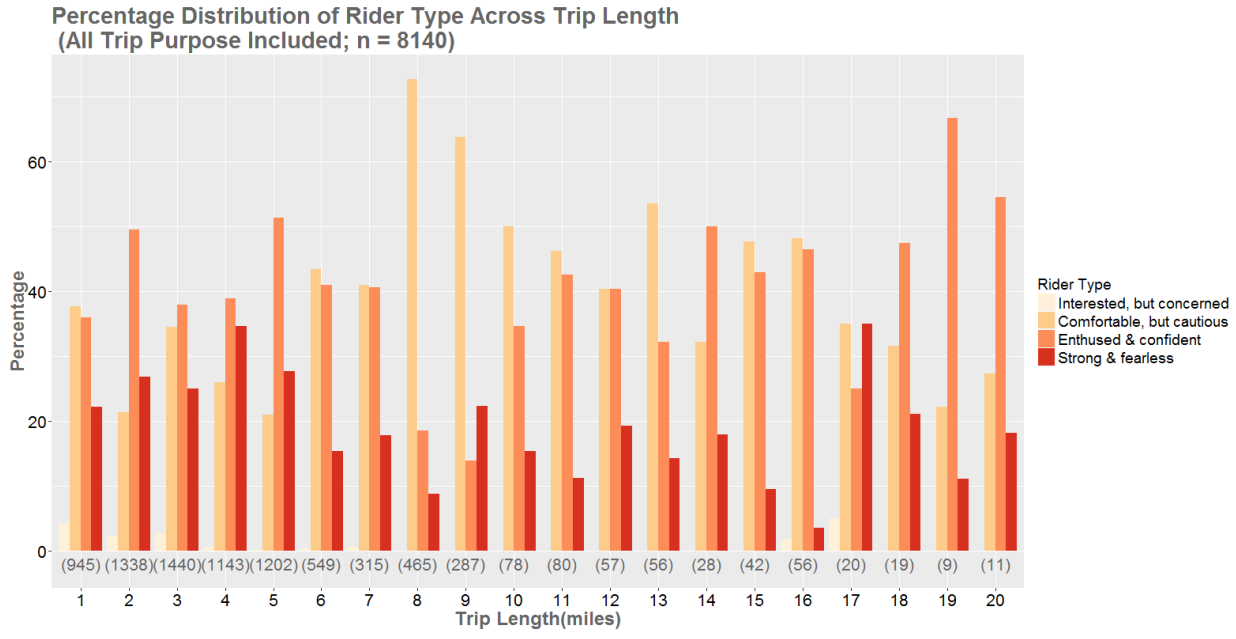


Figure 12(h)(i)(b) Rider Type Distribution across Trip Length for All Trip Purpose

Figure 12(a)-(i). Cycle Atlanta Trips Continued

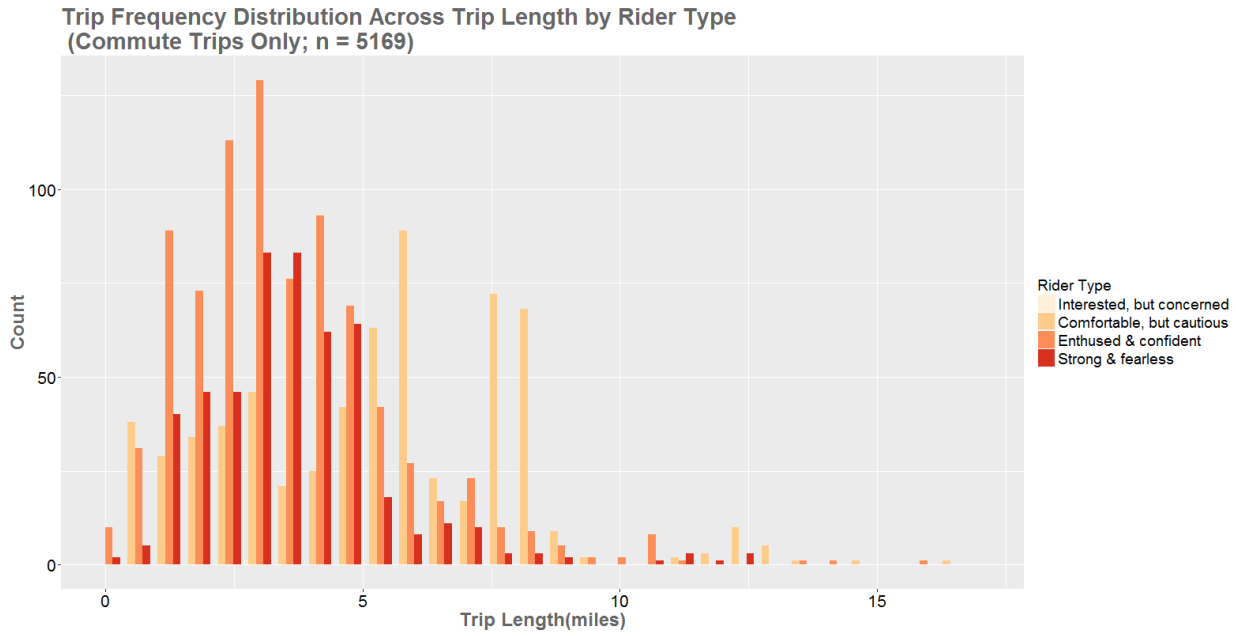


Figure 12(h)(ii)(a) Trip Frequency across Trip Length by Rider Type for Commute Trips

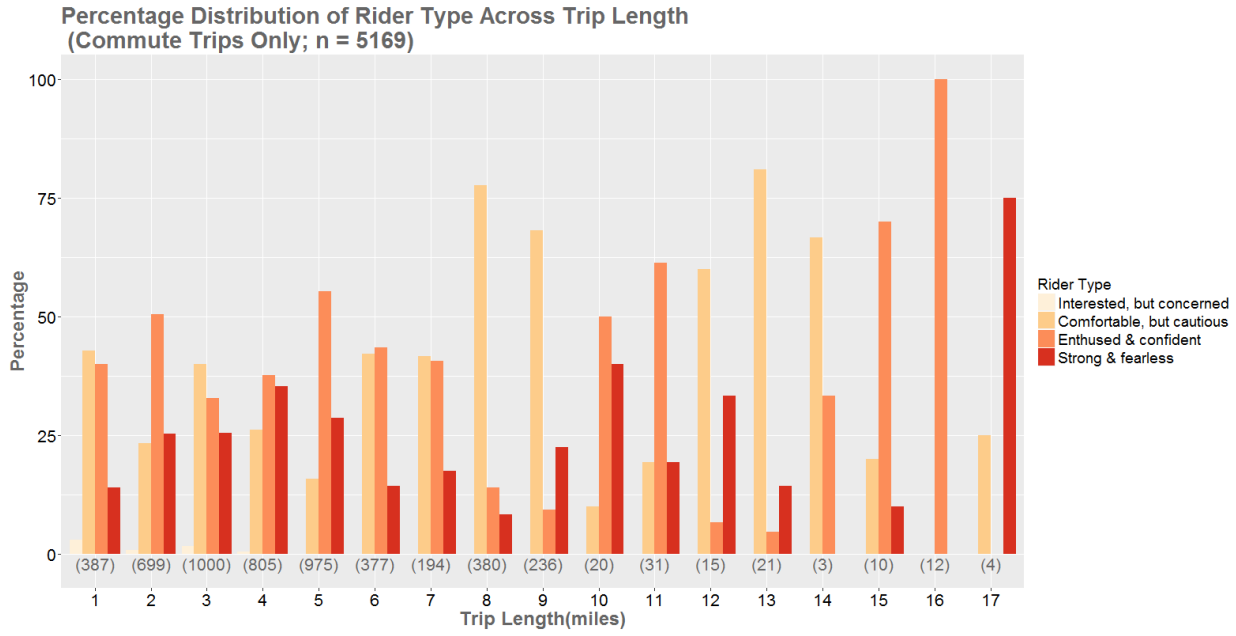


Figure 12(h)(ii)(b) Rider Type Distribution across Trip Length for Commute Trips

Figure 12(a)-(i). Cycle Atlanta Trips Continued

For the purpose of this research, the alternative to the chosen route was taken to be the shortest network distance path generated by the Dijkstra's algorithm in R. Relational models were constructed to understand (i) how rider characteristics influence trip length and (ii) the percent deviation of the chosen route from the predicted shortest route based on rider characteristics dependent upon a binary choice of whether the riders chose the shorter of the two routes. The majority of the trips (~60%) were found to be longer than the shortest network distance based path with mean deviation being 20% of the shortest network distance based path length and median deviation being close to 2% of the shortest path length.

Model Results

As mentioned earlier, two different analyses are presented in this chapter. The first analysis focuses on developing models to understand the relationship between trip length and age, gender, and rider types. These models are purely informational as trip length is determined by a lot of other factors than just socio demographics of riders. In addition, even between the same origins and destinations, some riders may choose a shorter route while some other riders may take a detour to access better facilities. These models serve more as a basis for future research to understand how trip length varies by rider types or other socio demographics even when the shortest network based distance for the trips are the same i.e., it provides the basis for segmentation of future route choice models into different groups. The second model, the main focus of this chapter, models the cyclists' decision to choose the shortest route or not and then, conditioned on that decision, their deviation from the shortest route based on road attributes and sociodemographic characteristics.

Trip Length Models

For all the models, the *interested but concerned* group was chosen as the base group, as was age 18-24 and gender male, implying that all results should be interpreted in comparison to that category. Different model types were estimated to obtain the best fit starting with linear regression which gave a model fit of 0.04 followed by a log transformation of the response variable trip length as (trip length + 1). This transformation was inspired by the distribution of trip length and also because a sufficient number of trips had trip length <1 mile. The log transformed model gave a model fit of 0.02. However, as mentioned before, in our dataset, trip length is only observed when trip length >0 making it an ideal candidate for truncated normal regression. So the next model fitted was a standard truncated normal regression model and the McFadden's $\rho 2$ obtained for that model (base equally likely) is 0.21 which is higher than both the linear and the log transformed model. Since for the shortest path and other route models, only trips that are longer than 1 mile were considered, we also ran a truncated normal and a censored normal regression with threshold at 1 mile. The truncated normal regression with truncation at trip length = 1 mile gives a lower model fit (0.13) than the censored regression with censoring at trip length = 1 mile (0.21). Additionally, the standard error across all the models is in the range of 2.4 – 3.4 which is quite high given that the mean trip length is 3.4 miles which partially accounts for low model fits. It should be noted though that as mentioned before, trip length is decided a lot more by other factors like time of day, trip purpose etc. than just socio-demographics and thus, the low model fit is not actually surprising.

Of particular interest however, is the sign change of coefficients that happen after adding the lower confidence groups in the model. With the *comfortable and cautious* group and the *interested but concerned* group as the base, both *strong and fearless* and *enthused and confident* group had negative coefficients indicating that their trip lengths are shorter than the other two groups which is also apparent from Figure 12. However, when compared to the *interested but concerned* group alone, all other groups have longer trip lengths. Table 10 presents the results of the regression model on trip length as function of sociodemographic characteristics of the riders. Age has a positive relationship with trip length and male riders are also more likely to ride longer distances. *Enthused and confident* riders are less likely to take longer trips than *comfortable but cautious* riders, and so are *strong and fearless* riders, although both these groups are more likely to travel longer distance than the *interested but concerned* group. This may be because *enthused and confident* and *strong and fearless* riders are more inclined to use shortest routes even if there are no bicycle facilities which renders their trip short compared to *comfortable but cautious* riders, but the interested but concerned riders in general ride only for short distances. The parameter estimates are fairly similar across the linear regression model and the censored normal regression model with linear regression estimates being identical to the censored regression models with censoring at 0 miles while the truncated model has slightly higher estimates than either of the two models.

Table 10. Trip Length as Function of Socio-demographic Characteristics

Exogenous Variables	Linear Regression	Log Transformed	Truncated Normal Regression (τ= 0)	Truncated Normal Regression (τ= 1)	Censored Normal Regression (τ= 0)	Censored Normal Regression (τ= 1)
	y= Trip Length (N=5163)	*y= log(Trip Length + 1) (N=5163)*	*y= Trip Length, trip length > 0 (N=5163)*	*y= Trip Length, trip length >1 (N=4830)*	*y= Trip Length, trip length>=0 (N=5163)*	*y= Trip Length, trip length>=1 (N=5163)*
	Estimate *(t-stat)*	Estimate *(t-stat)*	Estimate *(t-stat)*	Estimate *(t-stat)*	Estimate *(t-stat)*	Estimate *(t-stat)*
Intercept	0.868* *(2.117)*	0.899*** *(10.634)*	3.1 ** *(-3.258)*	-6.957*** *(-3.942)*	0.868 * *(2.118)*	1.315 *(1.783)*
Strong and Fearless	1.269** *(3.233)*	0.363*** *(4.489)*	3.363*** *(3.655)*	3.907* *(2.302)*	1.269 ** *(3.235)*	1.699*** *(3.939)*
Enthused and Confident	1.316*** *(3.39)*	0.344*** *(4.305)*	3.434*** *(3.75)*	4.417** *(2.609)*	1.316*** *(3.392)*	1.719*** *(4.026)*
Comfortable but cautious	2.131*** *(5.489)*	0.479*** *(5.987)*	4.668*** *(5.092)*	6.288*** *(3.704)*	2.131*** *(5.49)*	2.539*** *(5.942)*
Age	0.335*** *(4.18)*	0.061*** *(3.683)*	0.513*** *(4.143)*	0.593** *(3.11)*	0.335*** *(4.183)*	0.392*** *(4.578)*
Gender	0.454*** *(5.554)*	0.034* *(2.012)*	0.734*** *(5.528)*	1.668*** *(8.44)*	0.454*** *(5.558)*	0.394*** *(4.575)*
Sigma			3.00*** *(59.59)*	3.385*** *(42.237)*		
Logsigma					0.875*** *(89.058)*	0.923*** *(88.685)*
Model Statistics						
R2	0.04	0.023				
F- statistic	42.25	24.93				
Loglikelihood at EL			-14473.2	-11433.7	-15452.8	-14689.6
Loglikelihood at MS			-11640	-10195	-12086.7	-11852.95
Loglikelihood at Full Model			-11412	-9933.6	-11862	-11639.19
McFadden's ρ2(at EL)			0.212	0.13	0.232	0.208

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Shortest Route and Deviation from the Shortest Route

The second model was used to understand the relationship between deviations from the network based on the shortest route depending on route attributes and socio-demographics, particularly rider type. No previous study has related route preference to cyclist comfort and confidence level, therefore, this model may serve as a proxy to understand if any rider group is systematically choosing a longer route possibly because

of factors not yet known. One trip was randomly chosen for each user amongst the four trips in the dataset to perform the shortest route analysis. Since the data is significantly biased towards more confident riders, male and younger riders, one trip per user was chosen to keep the bias affecting model results at the minimum – in the future however, the analysis will be extended to include multiple dissimilar trips from each user and the results will be compared.

About 48% of the cyclists chose the network distance based shortest route for their trip. For the remaining 52% trips, the network based shortest route is shorter than the actual trip length which is similar as in full dataset implying that we have a fairly representative sample of the response variable. About 6% of the trips are almost twice the length of the network based shortest distance. Trip lengths in the reduced sample have a median of 3.36 miles which is slightly shorter than the full dataset median (3.6 miles) and a mean of 4 miles which is same as the original mean implying that we are slightly oversampling shorter trips in the reduced dataset. The mean deviation from the network based shortest path is ~20% while median deviation is about 4% which are similar to the corresponding values in the original dataset. For rider types, about 32% of the riders are in the category of *comfortable but cautious and interested, but concerned*, about 22% in the *strong and fearless* category and 46% in the category of *enthused and confident* which is similar to the original distribution of the rider types in the user dataset with slight oversampling of enthused and confident riders (originally 45%) and strong and fearless riders (originally 19%) and undersampling of comfortable but cautious and interested but concerned group (originally 36%). There are three reasons for this mismatch even when trips are randomly sampled – first, not all users who have provided

information on rider type have provided information on age and gender, so when we use a data set that is complete in information on all of these three variables, we lose some users. Second, not all users who provided sociodemographic information recorded trips – of the 1494 unique user ids, only 932 users recorded rider type, age and gender and of them, 821 users actually recorded trips – the remaining 111 users were lost while merging user data with trip data. Finally, by removing trips that are less than 1 mile, we may have introduced bias in favor of more confident riders who are more likely to take longer trips. However, since the bias is only about 1-2% in all categories, no corrective measure is applied, although we recognize the bias. The route attribute variables used for modelling are given in Table 1. In addition, we also used trip length as an explanatory variable with the idea that cyclists are less likely to deviate from the shortest route if it is already a long trip.

As mentioned before, two part model and sample selection model were used to estimate the deviation of a trip from the network based shortest route. The first part of both the models is a binary probit model that models the decision of the cyclist to choose the shortest route or not based on socio-demographic characteristics of the cyclist and shortest network distance, AADT, speed limit, % of trip having bicycle facility, and % of trip having sidewalk. In the case of the two part model that models the percentage deviation of the trip from its network based shortest distance, the second model is a logistic regression model. In the case of the sample selection model, the second model is a parametric truncated regression model, left truncated at zero, that models the percentage deviation and the two stages are related via the inverse Mills ratio as discussed in the literature review section. Mathematically,

$y^* = 0, \quad if\ shortest\ route\ chosen;$

$Else,$

$y = y^* = percent\ deviation\ from\ shortest\ path\ if\ shortest\ path\ not\ chosen$

Table 11 presents the results of the estimated models. As mentioned before, the variables used in the first part of the model are trip distance, AADT, speed limit, % of bicycle facility on the route and % of sidewalk on the route. Number of lanes and percentage of truck were also used in the initial stages of the modelling, but were found to be insignificant and hence removed from the final models. Age, gender and rider type were initially used as sociodemographic variables – however, other than age, none of the other two variables were significant. Additionally, gender and rider type are significantly correlated, so only rider type was retained in the final model. The exogenous variables in the second model were difference in AADT, speed limit, percent of bicycle facility and sidewalk between the chosen path and the shortest route along with trip length and rider type and age. For all the road attribute variables, the difference was calculated as *(value of the variable for the shortest route – value of the variable for the chosen route)* implying that any estimated coefficient with a positive sign will indicate that cyclists are likely to deviate more from the shortest route as the difference between the variable value at chosen route and at shortest route gets higher. Both age and rider type variables were converted into ordinal variables for the purpose of modelling with order increasing in the direction of higher age group and higher confidence level, respectively. It should be noted that in the binary model, since *choosing the shortest route is coded to be 0* and deviations from shortest route is actually what is being modelled, any estimate with a

positive sign indicate a cyclist's likelihood to deviate from the shortest route given higher values of the variable. It should also be noted that to maintain compatibility in the scale of the values for the variables used in the model, particularly between AADT, speed and bicycle facility percentage, AADT values are taken as AADT/1000 and speed values are taken as speed/10. Therefore, instead of interpreting results with respect to a unit change in AADT, results should actually be interpreted with respect to 1000 units change in AADT. Same method should also apply for speed.

Table 11. Deviation from Network based Shortest Route as Function of Road Attributes and Socio-Demographic Characteristics

	2 Part Model (2PM)				Sample Selection: 2 Step Estimator				Sample Selection: ML Estimation			
	Binary Probit (N = 437)		Fractional Logit Regression (N = 239)		Binary Probit (N = 437)		Regression (N = 239)		Selection Equation		Outcome Equation	
Variables	**Estimate**	***t-stat***	**Estimate**	***t-stat***	**Estimate**	***t-stat***	**Estimate**	***t-stat***	**Estimate**	***t-stat***	**Estimate**	***t-stat***
Intercept	**1.881*** **	*4.098*	-0.267	*-0.72*	**1.776.**	*1.699*	**0.486***	*2.558*	0.7954	0.718	**0.476*** **	*3.373*
Network based Shortest Distance	**-0.113*** **	*-4.13*	**-0.248*** **	*-6.13*	**-0.142 *** **	*-5.611*	**-0.0418***	*-2.297*	**-0.141*** **	*-5.586*	**-0.0426*** **	*-3.332*
AADT	**0.051*** **	*3.929*			**0.0186***	*1.987*			**0.0183.**	*1.932*		
Speed Limit	**0.035*** **	*8.281*			**0.065***	*2.364*			**0.065***	*2.388*		
% Bicycle Facility	**-1.236*** **	*-3.86*			**-0.404****	*-2.223*			**-0.4326****	*-2.279*		
% Sidewalk Facility	-0.034	*-0.12*			-0.030	*-1.136*			-0.030	*-1.143*		
Slope	**0.243****	*2.239*			**0.334***	*1.936*			**0.338.**	*-1.954*		
AADT difference			0.001	*0.079*			0.005	*0.091*			0.003	*0.064*
Speed Limit difference			0.004	*0.749*			0.001	*0.724*			0.001	*0.803*
% Bicycle Facility difference			**-0.64***	*-2.23*			**-0.465***	*-1.895*			**-0.455***	*-1.9*
% Sidewalk difference			-0.176	*-0.82*			-0.073	*-1.043*			-0.073	*-1.043*
Slope difference			**0.213***	*1.86*			**0.277***	*1.92*			**0.261***	*2.03*
Rider type	-0.043	*-0.66*	-0.004	*-0.06*	-0.01	*-0.058*	-0.003	*-0.019*	-0.013	*-0.221*	-0.001	*-0.033*
Age	**0.324****	*2.145*	**0.271***	*1.814*	**0.241.**	*1.735*	**0.152.**	*1.875*	**0.2414.**	*1.735*	**0.154.**	*1.655*
Model Statistics												
R2	**0.374**		**0.152**		**Multiple R2 = 0.15, Adjusted R2 = 0.11**				**LL(model) = -295.5588; LRT = 48.1 with 8 dof**			
Inverse Mills Ratio (λ)					**-0.051**				NA			
Sigma					**0.264**				**0.263**			
Rho (ρ)					**-0.192**				**-0.15**			

*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

It is interesting to note that the parameter estimates of all the three models are similar with no substantial difference in their estimates by three different methods. The estimate for rho (ρ), the correlation between the errors is relatively low, indicating that the errors are nearly independent, which may be one reason why the model results are so similar. The inverse Mills ratio is the additional term to be estimated in the two step sample selection model as compared to the two part model but that estimate is also insignificant (*t-stat = 0.937*) which may be another reason why estimates of two part model and two stage sample selection model are similar. The composite model fits are also similar, therefore, we provide a general discussion of the results with reference to the two part model that can easily be extended to the other two models.

The results from the first part of the models show that senior riders are more likely to not choose shortest route as are the low confidence riders. Models with gender as explanatory variable *(male = 0, female = 1)*, showed positive coefficients indicating that female riders are more likely to deviate from the shortest route than their male counterparts. Among the socio-demographic variables only age is significant while all the road attributes are significant except for percent of sidewalk on the route. Both AADT and speed limit have positive coefficient implying that higher the AADT and speed limit on the route, more likely that a cyclist will not choose the shortest route. On the other hand, percent of bicycle facility and sidewalk both have negative coefficient indicating the more there is bicycle facility or sidewalk, the cyclist is more likely to stay on the shortest route. Trip distance is significant and has a negative coefficient indicating that longer the trip length, a cyclist is less likely to deviate from the shortest route which is similar to the findings of Dill and McNeil (2012). Slope is also significant but with a

positive sign indicating that higher the altitude change on the route, the more likely that a cyclist will choose a detour. The model fit for the two part model is 0.37 which is reasonable but also indicates that there might be other street attributes like scenic beauty or parking that we did not use in this model and which can influence route choice decisions. From the odds ratio, calculated as exp(coefficient), an increase in 1000 units of AADT on the shortest route can have a 5% increase in the odds of choosing longer routes while an increase of 10 mph in speed limit increases the odds of deviating by 3.5%. A percent more of bicycle facility (for example 0.7% vs 1.7 %) on the shortest route decreases the odds of deviating by 70%.

The second model results also indicate that senior riders are likely to take longer detours than their younger counterparts and so are less confident riders and female riders. However, in the second model only trip distance, slope difference and age are significant. Trip distance has a negative sign indicating that people are less likely to make a detour for longer trips. Slope difference on the other hand has positive coefficient indicating that the higher the difference between altitude of the shortest route vs the chosen route, more deviation from the shortest route can be expected. Similarly, both AADT and speed limit difference come with a positive coefficient, although not significant, indicating that higher the difference between those variables on the shortest route versus on the chosen route, more deviation can be expected. Difference in percentage points in bicycle facility between the shortest and the chosen route comes with a negative sign indicating that higher the difference, the percentage deviation is less which alternatively can be interpreted as more percentage of bicycle facility on the shortest route implies less deviation from that route. An increase of 1000 vehicles in the difference between AADT

of the shortest route and the chosen route can increase deviation by 0.1% while an increase in the speed gradient can increase deviation by 0.3%. Presence of bicycle facility has the biggest contribution with ~60% less deviation when there is 1% less difference in percentage of bicycle facility between the shortest and the chosen path.

Discussion

The overarching theme of this chapter is to understand some of the trip making decisions of the cyclists of Atlanta who provided us data on their socio-demographics as well as on their route choices via the Cycle Atlanta app. The data are a subset of all cyclists in Atlanta, and in absence of any benchmark, it would be difficult to prove or disprove if the Cycle Atlanta users generally represented the cycling population of Atlanta in their trip making decisions. However, the results obtained in our analysis were fairly consistent with existing literature, common sense and user experience, and hence can be mostly generalized for the purpose of planning cycling infrastructure for Atlanta.

For all the models used in this chapter, a major drawback is the strong parametric assumptions required for the validity of all analysis to hold. Tobit censored and truncated regression models as well as two part models assume error terms to be homoskedastic and normally distributed and any violation of these assumptions renders the log likelihood formulation of these models wrong (Cameron and Trivedi 2005, Greene 2005). Therefore, it is important to check for these assumptions before any of these models are used for analysis.

The model fit for trip distance has been consistently low which indicates that there might be some variables other than socio-demographics that we are not including in

the analysis. Literature has shown trip purpose, time of day (morning commute vs evening commute), climate and other built environment related factors to influence trip making decision as well as trip length. Since we are only considering commute trips for this analysis, we could not use trip purpose as an exogenous variable. In future, this study can be extended by including time of day and climate as well as other route attributes. However, the models on deviation from shortest path provide some insight into factors that may influence trip length as well since longer trip lengths are more often longer deviations from the shortest path rather than longer 'trips'.

The results indicate that the decision to choose the shortest route or deviate predominantly depends on route attributes rather than rider type which implies that all cyclists, irrespective of their level of comfort or confidence prefer bicycle friendly streets with a significantly high elasticity for bicycle facility. This indicates that even if existing arterials cannot accommodate a separate facility, the planning agencies can invest in a facility away from the main arterial with a high possibility of cyclists actually using the facility. However, the mean detour that the cyclists take, according to our dataset, is about 30% of the shortest route and it is high compared to studies in other places, which ranges between 10 – 20%. A possible reason for the high mean detour in the routes recorded in Cycle Atlanta can be the extensive use of Atlanta Beltline trail which cyclists tend to go out of the way to access because it is the only really good protected facility that Atlanta cyclists have access to – cyclists in Atlanta are more restricted in their choice to deviate and how far to deviate than cyclists in Portland, Zurich or San Francisco and this higher deviation should be treated more as an extreme case than normal. A study of Cycle Atlanta trips by Mingus (2015) showed that a significant number of trips originate

from within the 2-3 mile bikeshed area around Beltline and these trips possibly would not have happened if the facility did not exist. Therefore, to encourage people to use bicycling as a mode of transportation although bicycle facilities need not be on main arterials, it should also be optimized enough to create a series connected facilities for cyclists that do not make the trip longer than ~ 20% of the shortest network distance.

Conclusion

In this chapter, we analyzed the Cycle Atlanta trips based on the socio-demographic characteristics of its users and route attributes. We also estimated the likelihood of a cyclist to choose a longer route over a network based shortest route depending on his/her socio-demographic makeup, confidence level, and route characteristics. The results show that female and senior cyclists are more likely to make commute trips that are shorter than their male and younger counterparts, and are more likely to take longer routes than the network based shortest routes. A major proportion of exercise trips were recorded by senior cyclists implying that they possibly use cycling for the purpose of exercise predominantly. However, it also implies that younger cyclists are taking up cycling as an alternative commute mode and will possibly be using bicycles to commute even at a later age instead of restricting its use for only exercise or recreation.

The only trip purpose that has comparable female and male cyclists even when the participant based is heavily biased towards male cyclists, is shopping. Therefore, to have more female cyclists, it may require a serious reconsideration of land use planning to create more accessible shopping destinations and to make trip chaining easier. The likelihood of deviating from the shortest route depends significantly on presence of cycling infrastructure and bicycle friendly street qualities like low AADT and speed

limit, irrespective of how confident or comfortable the cyclist may be. This indicates that instead of classifying cyclists as new or inexperienced cyclists and constructing infrastructure accordingly, it will be more effective if the connectivity of facilities in the locality is taken into consideration when planning for cycling infrastructure. Future research will look into incorporating other factors like traffic stress, scenery, and climate and time of day variables into the route choice model of cyclists in Atlanta.

CHAPTER 6

CHOICE SET GENERATION

Introduction

In recent times, bicycling as an alternative mode of transportation has been promoted both at the federal as well as local government level. However, the adoption of cycling as a mode of commute is considerably hampered by a perceived lack of safety that is often equated to a lack of infrastructure. An intuitive method to understand bicyclist's infrastructure preferences is to track their route preferences and collect data on their willingness to deviate from the shortest route to access designated facilities. Traditionally this has been done using stated preference surveys where bicyclists have been asked to state their preference for one treatment over another or factors that motivate or deter them from cycling. Recently, an alternative to stated preference surveys has gained popularity wherein data on vehicle trajectories are collected via vehicle tracking devices. Vehicle route choice preferences are then modeled based on revealed chosen trajectory.

Although frequently used for collecting data on automobiles, this approach has been rarely used for cyclists because of multiple issues. First, until recently, tracking devices were prohibitively costly and hence, were only used for cases of primary importance to travel demand modeling and traffic flow management. Second, computational effort required to model route choice is significant and its effectiveness is dependent on availability of a high resolution network. Often, street networks are not updated to include bicycle facilities recently constructed and bicyclists tend to use by-

lanes and cut-thrus that are rarely found in street networks, both of which render route choice modeling for bicyclists much less effective and much more complicated. Finally, bicyclists are much less likely to optimize routes based on travel time, which is the standard optimization algorithm used for vehicular traffic. Developing and using algorithms suited for modeling bicyclist route choice require separate efforts than the standard practice and agencies are often restricted by budget to allocate separate resources for cycling.

Of late, however, new computationally efficient algorithms have been designed and proposed to generate route alternatives that consider different optimization objectives like slope, scenery, traffic speed, and presence of facilities. Hood et al. (2011), Broach et al. (2012) and Menghini et al. (2011) have successfully used such algorithms on either standalone GPS device-based data or GPS enabled smartphone-based data to analyze and model route choices of cyclists in San Francisco, Portland, and Zurich, respectively.

However, availability of high resolution network data and matching GPS data to that network still remains an issue that hinders route choice modeling of cyclists. Additionally, literature suggests that perception of safety may depend on comfort and confidence level of cyclists, a spectrum which can be captured in the popular categorization of cyclists into four different categories of *strong and fearless*, *enthused and confident*, *interested but concerned*, and *no way no how* (Geller 2006). Therefore, generating a universal choice set for all cyclist types might result in poor estimation of the choice model and affect the prediction power of the model at a later stage. Some of the recently proposed algorithms attempt to capture behavioral variability by using a random draw of parameters from different distributions based on user type – however,

such algorithms are computationally expensive and thus, often require a compromise on the number of attributes that can be included in the cost function.

In this paper, we propose a choice set generation algorithm based on the theory of random utility maximization with the possibility of generating different choice sets for different rider types by using separate parametric forms of the unobserved random heterogeneity for different groups. The proposed algorithm is implemented on the network of Atlanta to generate a choice set for recorded cycling trips of the users of the Cycle Atlanta app. The generated alternatives are similar to that obtained by using labeling algorithms but are more varied than those obtained via k-shortest path algorithms and the method is less computationally intensive than existing stochastic methods of choice set generation.

Background and Motivation

There are three steps to the route choice problem: (1) path enumeration or generating a set of travel alternatives for any chosen route between an origin-destination pair, (2) estimating a disaggregate demand model based on individual route choices, and (3) predicting choice probabilities using the route choice model developed for different planning purposes. Route choice modelling presents unique challenges in all of the above categories. In this section, we present a brief contextual review of literature in the area of path enumeration and choice set generation.

Path Enumeration/Choice Set Generation

Multiple researchers have shown that size and composition of the generated choice set effects parameter estimates as well as prediction performance of route choice

models (Bekhor et al. 2006, Prato and Bekhor 2006, Freijinger 2007, and Bovy 2009). The size and composition of the choice set also affects the convergence rate and computational cost related to the choice set generation step. However, choice set generation is a complicated process as it is always a latent construct designed by the analyst: only the most preferred alternative is chosen and hence observed by the analyst in revealed preference route choice data. Theoretically, a region wide street network can provide an infinite set of alternatives for any chosen route between any origin-destination pair. However, the number of alternatives that are actually considered by the user is significantly constrained by feasibility (for example, cyclists cannot use highways and therefore, a possible route between origin destination that includes a highway link is not useful for cyclists), knowledge (users usually do not have complete information on the network they are using), and other idiosyncrasies that are particular to the user and the trip. Bovy and Stern (1990), Catalano (2003) and Hoogendoorn-Lanser (2005) define different categories of choice sets based on the type of alternatives contained in them. The *universal* set consists of all existing connections between the origin and destination pair including the infeasible and illogical ones (a path consisting of self - loops for example). At the first level of pruning, the *generated objective master set* is a subset of the universal set and is the set of paths that are logical; this is further subsetted as the *generated objective choice set* as the set consisting of routes that are feasible for the user to adopt. At the next level, Hoogendoorn-Lanser (2005) define *generated subjective choice set* which is a subset of the *generated objective choice set* and consist of routes that the user is aware of. The final choice set is the *generated considered set* which consists of alternatives that the user actually considers while making a decision and one

of such alternatives is finally chosen as the preferred route. According to Bovy and Stern (1990) and Catalano (2003), from a psychological perspective, users are capable of considering at most 7 alternatives while making a choice and thus, while having more alternatives provides a variety of choices, it might not represent the actual decision scenario.

There are two major approaches in choice set generation – the one that is used in this research explicitly considers choice set generation as a separate process than route choice and is done *prior* to route choice modelling. That is, the probability that a route is chosen is conditioned on the probability that the choice set is the actual choice set. Mathematically,

$$P(i) = \sum_{C \subseteq \Delta(M)} P(i|C)Q(C)$$

where C is a choice set in $\Delta(M)$, the set of subsets of M, $Q(C)$ is the probability that C is the true choice set, and $P(i|C)$ is the conditional probability of choice given set C (Swait 2001). This two-step formulation was first proposed by Manski (1977) and has been widely adopted in *apriori* choice set generation. The alternative to the two-step process is the iterative network assignment process where users are assigned to links/routes based on either user or system equilibrium over the network. No explicit choice set is enumerated in such cases and users are assigned to routes based on congestion or travel time optimization rather than any attitudinal preference or behavioral idiosyncrasies of the users. The biggest advantage of the two step method is that it provides the flexibility of incorporating behavioral aspects of users by explicitly including the choice set generation process as part of the route selection process. The two

step process is also computationally more tractable for large networks where the iterative process to generate optimal routes can be computationally intensive. However, a lot of the computational advantage is lost if the generated choice set is large and complex; in which case, solving the problem may become computationally infeasible. On the other hand, limiting choice set to any arbitrary number of alternatives for ease of computation leads to poor prediction performance for some model specifications (Horowitz and Louviere (1995), Bekhor and Prato (2006), Prato and Bekhor (2007), Bliemer and Bovy (2008)). Therefore, a trade-off is needed in determining the size of the choice set that sets an optimum balance between providing desired heterogeneity of choices and computational burden.

The computational efficiency in choice set generation and the quality of the alternatives generated (how closely they represent the actual behavioral/attitudinal preferences of the user while being sufficiently varied or well sampled from the set of all possible alternatives) also depends on the choice of search algorithm that can produce a reasonably competitive set of alternatives while being computationally efficient. Different path generation algorithms have been suggested and developed over time and this is still an ongoing field of research because of to the likely impact on both estimation and computation time and effort. Traditionally, generation of a set of alternatives to a chosen route has been based on shortest route algorithms (Prato 2009, Broach 2010). Several modifications and variations of the shortest route algorithm have been proposed in the literature over time and interested readers may find a summary of such advances in Prato (2009). Figure 13 shows the different path generation algorithms popularly used in route choice modeling loosely classified based on their method of computation.

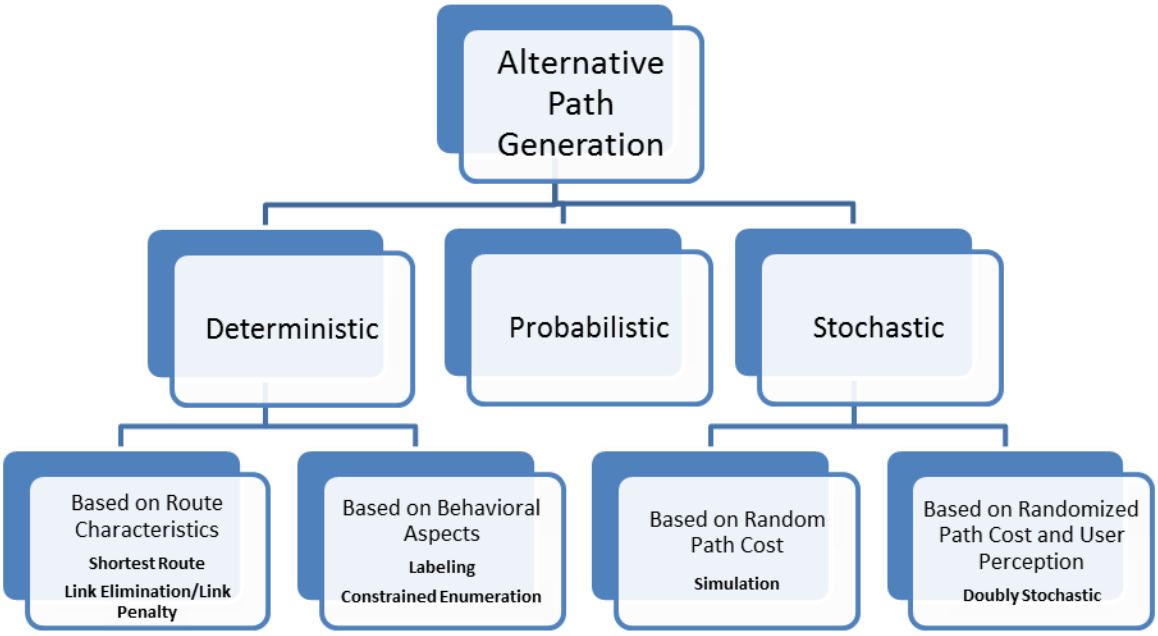


Figure 13. Different types of Path Generation Algorithms

From the behavioral aspect, Ben-Akiva et al. (1984) proposed a labeling algorithm where each label represented a particular route characteristic that a user might want to minimize or maximize – for example, a label may be the shortest distance or the least congested path. Ramming (2002) used the labeling algorithm to search for the shortest route based on 16 different labels while Prato and Bekhor (2006) used 4 attributes to generate alternatives. The results show that the success of the labeling algorithm is dependent on choice of labels, which is up to the discretion of the analyst and based on the understanding of the user preference and behavior.

Recent choice set generation literature shows incorporation of randomness of individual preferences and is based on importance sampling – the probability of choosing a route depends on the importance of route characteristics like distance or congestion. Prato (2009), Freijinger (2007), Bovy et al. (2009), and Frejinger et al. (2009) suggest

using a sampling correction term in the route choice model when choice alternatives are generated using these probabilistic methods.

In the simulation approach, it is assumed that users perceive path cost with some errors – so the cost function is assumed to be from a distribution and the results largely depend on the choice of the distribution from which the cost function is extracted. The doubly stochastic method is based on the assumption that both the path cost and the perception of the path cost vary among users – i.e. each user perceives path cost differently and with some error (Bovy 2009). The probabilistic methods are good at producing a heterogeneous set of alternatives and have been shown to replicate the observed route more frequently than other alternatives.

Two other methods not used very often in route choice are the constraint enumeration method and probabilistic method (Prato and Bekhor 2006, Bekhor and Prato 2009, Friedrich et al. 2001). The constrained enumeration method uses branch and bound algorithms and is based on the idea that instead of least cost, users often choose based on personal preferences. The algorithm was found to replicate the observed route closely but the computation time increases exponentially with the depth of the branching tree which is significant for a large network and is therefore limited in scope to small networks. The probabilistic method (Frejinger et al. 2009) assigns a generated probability to each route and calculates the probability of a route to be chosen conditioned on that generated probability. The method is therefore again computationally prohibitive to be carried out for a large regional network.

Choice Set Generation in the Context of Cyclist Route Choice

Choice set generation for cyclists is particularly complex on multiple levels. First, as already mentioned, the objectives of route optimization for cyclists are quite different than that for motorists. Second, cyclists extensively use short cuts through parking lots and through parks and other similar facilities that are normally not part of street network. Third, cyclists often use sidewalks to ride in the opposite direction, though not legally permitted to do so – it is difficult to identify a choice set for such behavior. Fourth, route choice of cyclists also depends on the purpose of the trip – exercise trips are significantly longer than commute trips and are often intentionally chosen to be more onerous than commute trips. Finally, since infrastructure preference of cyclists depend on the comfort and confidence level of cyclists, it may be postulated that the competing alternatives for a group of cyclists who are less confident, are going to be different from that of the cyclists who are confident and more comfortable.

Of the bicyclist route choice studies done to date, five studies are relevant and related to this research. The first two of those studies are stated preference surveys where the participants were given pre-determined routes to choose from (Aultman-Hall et al. 1998, Stinson and Bhat 2008). Aultman-Hall et al. (1998) provided the participants with only two alternatives, where one of the alternative was the shortest network distance path between origin and destination. Stinson and Bhat (2008), however, provided the participants with multiple alternatives to choose from. In both the studies, the alternative/s were constructed based on trade-offs among different attributes (distance vs slope and/or traffic speed/ traffic volume) and the participants were asked to choose one alternative that they found preferable. Stinson and Bhat (2008) also used a different set of

alternatives for different participants, although the variation in choice set was done randomly rather than basing it off participant characteristics.

Studies by Hood et al. (2009), Menghini et al. (2010) and Broach et al. (2012) used GPS based revealed preference route data for modelling route choice of cyclists in San Francisco, Zurich and Portland, respectively. Hood et al. (2009) used a doubly stochastic method (Bovy and Fiorenzo-Catalano, 2007) for generating the choice set where both the co-efficients of the attributes and the link cost were randomly simulated from chosen distributions. The doubly stochastic method is based on an iterative implementation of Dijkstra's algorithm where at each run, link costs are randomly drawn from one distribution and parameters are randomly drawn from another distribution. The literature suggests that even though the parameters and the link costs are chosen from two distributions, often they return identical routes and hence, only unique routes are preserved at each iteration of the algorithm and the path search is continued until the desired number of alternatives has been generated. The doubly stochastic method has been used with mostly with single attribute cost function because of computational effort, however Hood et al. (2011) reported using both route length and travel time as attributes for their study and having better performance than breadth first search algorithm (BFS).

Menghini et al. (2010) used breadth first search (BFS) with link elimination in MATSim holding route length as the constraint. This method is also based on the shortest path search in which one link at a time is systematically removed from that shortest path to generate the next shortest path connecting the origin and destination in the new network. The algorithm stops once the desired number of alternatives has been

found or when there is no feasible connected route remaining between the origin and destination.

Broach et al. (2012) used a modified version of the original labeling algorithm where they calibrated the parameters of the labels using the metric of deviation of the generated routes from the corresponding shortest route as compared to the deviation of the observed route from the corresponding shortest route. They compared the performance of their algorithm with k-shortest path, simulated link cost and labeling algorithms on the same dataset and found that the calibrated labeling algorithms provided a greater variability in number of alternatives in choice sets and also replicated the observed route more frequently than other algorithms. The results were, however, later disputed by Halldórsdóttir et al. (2014) who could not replicate them for their own study.

Halldórsdóttir et al. (2014) compared the performance of the three most commonly used choice set generation algorithms for cycling trips, namely, doubly stochastic, breadth first search with link elimination, and branch and bound. They used more specific bicycling related attributes in the cost function when generating the choice set and found that the doubly stochastic algorithm provides better coverage and replicated the observed route in more cases than either of the other two algorithms. However, doubly stochastic is computationally expensive and the breadth first search performs better in that area. Branch and bound algorithm were found to perform worse than either of the other two algorithms on quality of choice set and computation effort.

Quality of Choice Set

Halldórsdóttir et al. (2014) suggested a number of ways to evaluate the effectiveness of the choice set generation algorithm as well as to measure the quality of the choice set generated. The first and foremost measure of effectiveness and consistency for a choice set generation algorithm is whether it replicated the observed behavior, i.e., the observed chosen route. Halldórsdóttir et al. (2014) extended this measure to evaluate the coverage of the generated choice set. They defined a parameter $\boldsymbol{O_{nr}} = \frac{\boldsymbol{L_{nr}}}{\boldsymbol{L_n}}$ as overlap measure where $\boldsymbol{L_{nr}}$ is the overlap length between generated path and chosen path and $\boldsymbol{L_n}$ is the length of the chosen path. They then define coverage as "percentage of observations for which an algorithm generates a route that satisfies a particular threshold for the overlap measure: $\boldsymbol{max_r} \sum_{\boldsymbol{n=1}}^{\boldsymbol{N}} \boldsymbol{I}(\boldsymbol{O_{nr}} \geq \boldsymbol{\delta})$" where $\boldsymbol{I}(.)$ is the identity coverage function and $\boldsymbol{\delta}$ is the threshold for overlap measure. Halldórsdóttir et al. (2014) also suggests using the path size correction factor given by Ben-Akiva and Bierlaire (1999) as a measure of the heterogeneity of choice set generated where the path size factor takes values between 0 and 1 with 0 indicating identical route and 1 indicating unique routes. They further used the consistency index developed by Bekhor and Prato (2006) to measure the behavioral consistency by aggregating the overlap measure generated earlier over all paths generated and comparing it to 100% overlap across all routes.

Choice Set Generation Methodology

This research presumes that having a distinct choice set generation step is methodologically superior to iterative assignment methods, because it provides the required flexibility that is needed to include behavioral differences among different user groups (Prato 2009, Fioranzo-Catalano and Bovy 2007). Unlike vehicular traffic, cyclist

route choice is often predominantly determined by safety and comfort concerns, which are subjective and perception dependent. Therefore, in modelling cyclist route choice, it is important that the methodology used can account for user taste heterogeneity as cyclists vary widely in their perception of safety and hence, in their preference for route infrastructure. Explicit choice set generation provides the required flexibility where the alternatives to the chosen route can be designed to be weighted differently on different attributes based on user characteristics and comfort with the said attribute.

Like most studies in the literature, the choice set generation methodology used in this research is based on the concept that the cost function associated with a route can be assumed to be the utility that the user associates with that route. This utility, in turn, is assumed to be additive/ additive-subtractive linear in parameter function of road attributes. Mathematically,

$$Cost\ C(x)$$

$$= U_{ix}$$

$$= \alpha_{1x}(road\ attribute1)_{ix} + \alpha_{2x}(road\ attribute2)_{ix} + \cdots + \alpha_{Nx}(road\ attribute\ N)_{ix} + \varepsilon_{ix}$$

where U_{ix} is the utility of the route x as perceived by user i, α_{1x}... α_{Nx} denote the vector of parameters that can represent the weight that a user attaches to a particular road attribute, and ε_{ix} is the random term that can account for factors not observed by the researcher. As mentioned previously, stochastic and doubly stochastic algorithms of choice set generation rely heavily on a random draw of the parameters from different distributions for different user types while labelling algorithms are more concerned with an appropriate combination of the road attributes that a user will realistically consider.

However, while a combination of labeling algorithm and doubly stochastic algorithm provides the most realistic behavioral foundation for choice set generation where the user group is significantly heterogeneous, it involves significant computational effort and is almost intractable in case of high resolution network and multiple road attributes (Halldórsdóttir et al. 2014).

In this research, we propose an alternative, less computationally intensive method of the choice set generation that is still capable of capturing the possible difference in preferences of different user types. We also use results of parameter estimates from the later step of route choice modelling to calibrate the choice set generated at the previous step and form an iterative process whereby we ensure that our choice set formulation most closely resembles a real life choice scenario for any future estimation process. Further, none of the previous research has experimented with different parametric forms of the error term. Since the distribution of the difference in random error terms among alternatives determines the probability of choosing an alternative, the rate of change of that probability is determined by the slope of the cumulative distribution curve, which, in turn, is determined by the shape and location parameter of the distribution. Now, if it is postulated that a less confident user has a higher rate of attrition for the same amount of increase in traffic speed or traffic volume than a more confident rider, then along with a deterministic part of the utility, we can use distributions with different shape and location parameters for different user groups to account for unobserved taste heterogeneity and thus generate different sets of alternatives for different user types.

Our choice set generation method starts from a previous model estimate on choice of shortest route on the same dataset. In our earlier research, we estimated a binary

logistic regression model regarding whether a user will choose the shortest network distance based route or not based on different road attributes. In this research, we use the parameter estimates from that model to construct the cost function for the choice set generation and introduce a Gumbel distributed random term in the cost function to account for the variability in preferences among different user types. The Gumbel distributed term is designed to have variable shape and location parameter for different users based on the sociodemographic variable rider type. The steps of generating choice set based on the method given are as follows:

- Step 1: Construct a link cost function

$Cost\ C(x)$

- $= \alpha_{1x}(road\ attribute1)_{ix} + \alpha_{2x}(road\ attribute2)_{ix} + \cdots + \alpha_{Nx}(road\ attribute\ N)_{ix} + \varepsilon_{ix}$ where α_{1x}... α_{Nx} are the co-efficient estimates from the shortest path model

- Step 2: Generate ε_{ix} as random draws from a Gumbel distribution with specified shape and location parameter.

- Step 3: Generate 10 – shortest paths using *k*-shortest path algorithm with the link cost calculated in Step 1 and Step 2 as the weight for each link.

- Step 4: Check if the path overlap of any generated route is more than 80% or less than 20% using path size overlap correction term (Casetta 2001), then discard that route.

- Step 5: Run path-size logit model on the generated choice set of (10 shortest paths + the path chosen) with road attributes as explanatory variables
- Step 6: Check if the estimated co-efficients are within 30% of the co-efficient estimates of the shortest path model
- Iterate until the co-efficients from the two processes are within threshold

The greatest advantage of this method is that it is semi-deterministic with only one random term to be generated which makes it computationally much more tractable. Also because of its semi deterministic nature, the cost function can include multiple attributes which are found to be significant determinants of route choice decision of cyclists from previous research, without an exponential increase in computational cost. The random term adds the behavioral indeterminacy or taste heterogeneity that we expect to see among users from different confidence and comfort levels. In cases where the user type is not known a priori, we may generate a varied choice set with samples from all the different distributions that we presume may represent the rate of adoption/attrition for each user group. The path size overlap check at the next step can ensure that generated alternatives are consistent with the user's choice by removing paths that do not have any overlap between them. The path size thresholds, the choice set size and the convergence thresholds are determined based on our experience from the shortest path model where the mean deviation from the shortest path has been about 30% and of the 40% of routes that deviated from the shortest path, none of them were completely unique with maximum overlap with the shortest path being 90%.

Calibrating the generated choice set with co-efficient estimates from the choice modelling step can also potentially help in addressing the user heterogeneity aspect of choice set generation. If the parameter estimates from choice modelling are found to differ significantly across the rider types or age or gender, separate cost functions for different groups can be designed to accommodate that variance across groups and that can help improve the quality of the generated choice set for the second round of iteration.

Results

The average overlap was measured by taking mean of the coverage parameter $\boldsymbol{O_{nr}}$ defined earlier in this chapter as $\boldsymbol{O_{nr}} = \frac{\boldsymbol{L_{nr}}}{\boldsymbol{L_n}}$ where $\frac{\boldsymbol{L_{nr}}}{\boldsymbol{L_n}}$ is the ratio of overlapping link length to length of the chosen route (Halldórsdóttir et al. 2014). The mean of $\boldsymbol{O_{nr}}$ is calculated for each set and the modal value of those mean values is taken as the average overlap of generated choice set. According to this measure, the average overlap is about 45% although shorter routes (trip length < 5 miles) had more overlap than longer routes. The method failed to find an alternative to all trips with length < 1 mile with the threshold to discard alternatives with more than 80% overlap and generated trip cost ≤ 2 × chosen trip cost. The method also failed to find more than 2 alternatives to the trips that are over 10 miles and found no alternatives for trips > 13 miles. This may be due to the fact that the network used, though of high resolution, is clipped at a 5-mile radius of a chosen point, effectively reducing the possible infinite set of alternatives to any chosen route. It should be noted though that irrespective of using a clipped network or not, most algorithms report similar problems for longer routes.

Discussion

Because of the computational effort needed in stochastic methods for high resolution networks like that of Atlanta, we focused on deterministic methods of choice set generation for this research. However, since previous research has shown that route and infrastructure perception may vary across cyclist types, it was also necessary to address the behavioral aspect of choice set generation along with its computational efficiency. Two methods were explored - first, adding a random term to the cost estimate to account for taste heterogeneity, and second, generating segmented choice sets for different user groups when the choice estimation step clearly shows a significant difference in parameter estimates across different groups.

The addition of the random term did not significantly alter the path search or the choice set generated, however, estimated path size logit models at the choice estimation stage did show significant differences across age groups. Therefore, three different cost functions were designed and only the relevant one was used in the second iteration step depending on the age of the user. In no case did it need more than two iterations to converge within the stipulated threshold.

The biggest disadvantage of the k –shortest path algorithm is that the generated routes are very similar to each other which when applied to choice set generation implies that the generated alternatives may lack in sufficient variability. On the other hand, the proponents of using k-shortest path for route generation point to the fact that people generally consider routes that are more or less similar or have large overlap – it is unlikely that a realistic choice set will consist of all unique routes without overlap. However, in case there is more overlap than is acceptable to maintain the range of

variation desired among the alternatives, an overlap threshold can always be specified which can help in discarding trips that are very similar to the chosen route.

An alternative deterministic algorithm to the k-shortest path would have been the labeling algorithm which is grounded on the philosophy that people first limit their choices based on one non-compensatory criteria, which happen to be either trip distance or travel time. Once the choice set has been narrowed down by routes comparable on the selected criteria, people add other criteria that are important to them to the original criteria and make a decision. At this stage, the choice may be compensatory or non-compensatory, but in no case is the choice made without a consideration of the original selection criteria. This algorithm is particularly helpful for generating cycling routes as different users have different preferences – while some may want to maximize bicycling on designated facilities, some may want to use routes with scenic beauty. Using this algorithm, it is easier to generate routes that maximize any particular label instead of a composite set of route attributes. This also can theoretically provide a greater variety of alternatives as routes that maximize one label can be completely different than another that maximizes a different label. However, the issues with labeling algorithm are (1) to take advantage of its ability to generate a wide range of alternatives, an equally varied range of labels have to be created, (2) more often than not, the labeling algorithm fails to find a sufficient number of alternatives based on the label it maximizes and (3) the quality of the choice set generated is highly dependent on the labels created which in turn depend on the perception and experience of the analyst. Additionally, critics point out that realistically, people do not consider a widely varying range of alternatives when making a trip and tend to deviate minimally from the shortest path. Therefore, proper

understanding is needed to balance the optimization between shortest distance and any other road attribute which depends on the discretion of the analyst.

CHAPTER 7

ROUTE CHOICE MODELLING

Introduction

The bicyclist route choice study by Aultman-Hall et al. (1997) was followed up by a series of such studies – these studies overcame the data issues of the previous study by using a GPS-based automatic route data collection system which minimized user induced bias in the data, had the advantage of recording multiple trips for each individual, and also significantly increased the number of participants as now the participant either only had to carry an instrument that recorded the trip automatically or had to download and turn on a smartphone app to record his/her trip.

Two parallel data collection methods emerged using GPS-based route data for bicycle route choice modeling around the same time. In the first case, GPS enabled devices were distributed among the participants at the beginning of the study for a scheduled period of time, and the participant either had to carry it with him/her or had to mount the device on his/her bicycle. The data recorded were locally stored in the memory card of the device and at the end of the scheduled period, the devices were called back from the participants and the data were retrieved and analyzed. Proponents of this approach include the Oregon bicyclee research group (Broach et al. 2009, Dill et al. 2008) and the travel demand modeling group headed by Dr. Kay Axhausen at Eidgenossische Technische Hochschule, Zurich (Menghini et al. 2010) who have used this approach for data collection and route choice analysis.

The other method of data collection is to use the GPS enabled smartphones owned by participants for recording data and directly uploading the data wirelessly from the phone to a central server. The advantage in this method is that it is cheap and does not require any investment in equipment while data quality is comparable to the GPS devices. Pioneering work in this area was done at San Francisco County Transportation Authority (SFCTA) (Charlton et al. 2009, Hood et al. 2009) and was later extended by the Cycle Atlanta group (Poznanski 2013). Hudson et al. (2012) also did a similar route choice study in Austin.

In the following section, the research using these two different technologies is discussed in detail, and a critique of the models developed in these studies is also presented.

GPS Device Based Data Collection Efforts

Dill et al. (2008) and Broach et al. (2009) both based their research on the same study of 162 cyclists in Portland, Oregon. While Dill et al. (2008) used the collected data to compare chosen route against the shortest path, Broach et al. (2009) extended the study to develop a path choice model for cyclists in Portland, Oregon. The studies included both utilitarian and recreational trips and participants were chosen through stratified sampling from respondents to an online survey. The demographic and personal characteristics used for stratification were cycling frequency (frequent vs. infrequent), home location (Portland vs. remainder), age, and gender. As mentioned earlier, at the beginning of their study, GPS devices were distributed to the participants, and at the beginning of each trip, the participant had to tap the screen to turn the device on. The user also had to enter a few other pieces of information by choosing from drop down menus

provided: including the trip destination category (home, school, work, etc.), the weather (sunny, cloudy, rain, etc.), temperature (hot, moderate, etc.) and wind (heavy, light, or no wind). Enroute, the GPS unit recorded location data at every 3 seconds, and these data were stored in the device. At the end of the study period, the data were retrieved and each individual trip was mapped. Participants were then asked to log on to their maps and identify any trip that was recorded erroneously – this was supplemented by a questionnaire to validate the correctness of the data collected and to understand the reason behind the choice to bicycle.

Similar to the study by Hall et al. (1997), in this case also, snapping the chosen route to the city network required augmentation of the network with links from the route data collected and validation through aerial photography. Map matching algorithms were developed by the researchers to account for GPS related errors like snapping onto adjacent roads instead of the actual route, data point clouds at intersections, and at start and stop, erroneous turns, etc. The final cleaned and matched network data were used for the analysis.

A statistical analysis of the data revealed that women traveled less distance than men and also had a higher mean rating for most factors influencing route choice as compared to men – significantly different mean ratings were observed between men and women bicyclists for minimizing distance, avoiding streets with high traffic and slopes. However, men had a slightly higher mean rating than women for riding on bicycle lanes. The most important factor in choosing a route was stated to be minimum time followed by low traffic volume and presence of a bike lane. No significant relationship was found between route choice and slope. A comparison between shortest route and the actual

route showed that people spent more time on bicycle facilities and low traffic streets than predicted by the shortest route and that the deviation from shortest route increased with length of trip.

Broach et al. extended the study by Dill et al. (2008) to develop a multiple variable discrete choice model of bike route choice of cyclists in Portland. In doing so, Broach et al. overcame the issue of comparing actual route choice with only the shortest distance – the model now was capable of predicting marginal utilities of different attributes and handling any interaction between them. One of the difficulties in developing a discrete choice model is generating a feasible set of not-chosen alternatives. In a city road network, this has a very large number of possibilities and hence, a sorting algorithm needs to be used to restrict the choice set to a manageable number. In this case, a labeling algorithm was used after trial and error with a few other algorithms. On average, 20 different alternatives were generated for every route. For the choice model, a multinomial logit model was used with correction for path choice overlapping using path size correction term from Ben-Akiva and Bierlaire (1999). The path attributes used for the model were distance, slope, turns, traffic volume, signals, and bike facility type. With all other parameters held constant, log distance was the most important factor in route choice, implying that for a short commute, a cyclist will be less willing to take the same detour than he/she would be if the commute was longer. Slopes and turns consistently had negative coefficients implying a disincentive attached to routes with high slopes or a significant number of turns. Traffic volume also proved to be a disutility, while traffic signals had a positive utility when the cross traffic was high and had a disutility for low traffic streets. Bike boulevards and paths were strongly preferred while the utility

associated with bike lanes was just enough to offset the disutility of traffic volume in that link. Therefore, bike lanes are preferred in streets with high traffic and over busy arterials without any bike lane. The route choice model developed by Broach et al. is being incorporated into the regional travel demand model of Metro, the Portland area municipal planning organization (MPO) in an effort to better predict where cyclists travel and what type of facilities they prefer, so that optimized investment decisions can be made.

Menghini et al. (2010) did a similar study on bicyclists in Zurich – however, they did not directly conduct the data collection for the study but rather received a multimodal travel dataset from a private agency in Zurich. The unique contribution of this study is in developing a GPS data cleaning algorithm for large datasets without any other information. The choice set was generated under MATSim (Multi-agent Transportation Simulation), using a search and bound algorithm, which generated about 60 alternatives to each route. The route choice model selected is the multinomial logit model with path choice overlap correction as in the study by Broach et al. (2009). The parameters used for estimation are maximum and average gradient, length of trip, percentage of marked bike facility, and number of traffic lights. Length was found to be the single most important criterion followed by average gradient and percentage of marked bike facility. Number of traffic lights and maximum gradient did not have any significant impact on route choice.

Limitations of the Studies

While the study by Broach et al. (2009) was one of the first to use revealed preference data for creating a route choice model for cyclists, a few limitations still remain that require further research. For instance, the study was based in Portland, a very

bike friendly city with more bike facilities than can be expected in an average U.S. city. Therefore, it is difficult to translate to other regions the findings regarding preference of bike facility and the willingness to travel an extra distance to avail oneself of a facility. Second, the GPS data cleaning was done manually, which was possible because of a low participant base and a low number of trips (1559 trips) used in the study. For larger datasets, it will not be possible to clean GPS data manually, and some algorithms and scripts will be necessary. But the most important issue with the studies by Dill et al. (2008) and Broach et al. (2009) is similar to that noted in the study by Hall et al. (1997) - all these studies lack segmentation of the cyclist population based on experience, comfort level, or attitude, although the literature has always emphasized the impact of rider characteristics on route choice decisions (Pucher and Buehler 2007, Krizek 2007). During data collection, cyclists were categorized as frequent and infrequent cyclists but no separate analysis was performed, possibly because the number of infrequent cyclists was very low.

Menghini et al. (2010) acknowledge this limitation of their study and suggest using socio-demographic characteristics like age, gender, and a measure of risk aversion of the riders to overcome the issue. As the remaining part of the literature critique will show, although age and gender have been used in one of the models, a measure of risk aversion and the level of experience of the cyclist still remain to be modeled into route choice studies.

GPS Enabled Smartphone Based Data Collection Effort

The data collection effort of this group used a free smartphone app called CycleTracks that was created and developed at San Francisco County Transportation

Authority (SFCTA). A user has to only download the app and turn it on at the start of the trip - the app then records route GPS data for every second of travel and stores it locally. On completion of the trip, the user is given the option to upload the trip or discard it. Upon uploading, the trip data are stored in a central server before it can be used for analysis (Charlton et al. 2009). As the goal of the study was to relate cyclist route choice to personal, trip-based, and network characteristic-based factors, the app comes with an optional provision where participants can provide their age, gender, cycling frequency, and the purpose of the trip. There is also a provision to enter the participant's email address, should he/she choose to do so. For maintaining anonymity of data, this field is completely scrubbed off during data analysis and is only stored for the purpose of future correspondence or survey needs.

The purpose of the CycleTracks research was to develop a bicycle route choice module for the existing tour-based travel demand model SF-CHAMP for San Francisco and the Bay Area. The data collected through CycleTracks were used to develop a multinomial logit model for cyclist path choice from which logsums are fed into SF-CHAMP, enabling it to assign the generated trips to the city road network. Hood et al. (2009) used the data cleaning and map matching algorithm developed at ETH by Dr. Kay Axhausen and Dr. Nadine Schussler (Schussler and Axhausen 2009) and used the same multinomial logit model with correction for path size overlap as was done by Broach et al. (2009) and Menghini et al. (2010). The contributions of this study are in using a different algorithm for the choice set generation, including gender and cycling frequency as model parameters, using panel data for model validation, and in extending the modeling exercise into a benefit cost analysis of possible new facility construction. The

choice set generation algorithm used is a doubly stochastic genetic search algorithm that generates the choice set through randomizing both the link attributes and the beta coefficients of the cost function. The parameters that were estimated in the model are for length, number of turns per km, upslope, type of bike facility, and travel in the wrong direction. The results suggest that cyclists prefer shorter routes and fewer turns, whereas upslope is always a disutility. However, bike lanes were found to be preferred over shared use lanes, and infrequent bicyclists were more likely to prefer a bike lane than shared lane. Slopes were particularly avoided by female cyclists and during a commute trip. A holdback sample of 202 cyclists was used to validate the results of the model.

The CycleTracks app was adopted for a similar study in Austin, Texas by Hudson et al. (2012). The data collected using the app were used to develop a cycling route choice map for the region, but was not extended to modeling route preferences.

Limitations of the Studies

Although the study by Hood et al. (2009) addressed most of the issues discussed previously, it still falls short of including a risk aversion attitude measure into the route model. Another issue in this study was the unequal numbers of trips recorded by users. To solve the problem, in the log likelihood function, each observation was weighted by the inverse of the number of observations, so that each individual had equal weight in the model. However, this raises the question of whether a cyclist who used a route once should have equal importance in model estimation as someone who uses the route regularly. Further investigation is needed to address such questions. Also, the benefit-cost analysis done in this study is based on a national cost estimate and calculated the user benefit or willingness to pay only based on value of time while several other factors

like health benefits and environmental benefits are more important for bicyclists. Therefore, further research is needed in developing a benefit cost analysis framework for cycling facilities that include all relevant factors.

Overall from the literature review it can be seen that all the route choice models for bicyclists are performed without any particular attention to difference in trip making pattern among cyclists of different socio-demographic makeup and/or attitudinal selectiveness. This research makes a significant contribution in that area by (i) including attitudinal variables rider type and socio-demographic variables like age and gender into the route choice model to understand how these variables influence route choice decisions and (ii) by designing segmented models for different age groups, gender and rider type – due to the inherent bias on male, younger and more confident riders in the dataset, any model estimated on the entire dataset has a possibility to mask the preferences and choices of the lesser represented groups. Segmented models provide the opportunity to separate out groups that are homogeneous in their preferences and choices from the ones that are significantly different.

Methodology

The primary data source for this research is the revealed preference trip data obtained from users of the GPS enabled smartphone application (app) Cycle Atlanta (Misra et al. 2014). The users of the app can voluntarily provide their socio-demographic, riding history, cycling frequency and trip purpose information, while the app, once switched on, collects the route trace of the bicycling trips they make at one GPS point per second. Once a trip is completed, upon receiving permission from the user, the GPS data are uploaded to a secure datacenter at Georgia Tech. Three separate databases are created

from the data received from Cycle Atlanta users: (i) users - which stores user related information like socio-demographics, rider history and cycling frequency, uniquely tied to user ids; (ii) trips – which stores information on trip purpose, start and end GPS points, duration of the trip and any other input on the trips from the users, uniquely identified by trip id and user id; and (iii) routes – which stores the complete GPS information for every trip including latitude-longitude of each point, timestamp, accuracy measures, speed and altitude, uniquely tied to trip id. Between October 2012 and June 2014, Cycle Atlanta had about 20,000 trips recorded by 1495 users, 60% of whom provided sociodemographic and cycling related information.

Variables

Two different sets of variables were used in these models – infrastructure related variables and sociodemographic variables. Table 1 shows the link attribute variables, their weighting factors and their route level aggregation methods that were used for this analysis. The principal idea behind using weighted averages of variables was to understand the effect of each attribute as a function of the link length. Except for the path overlap correction factor, all attributes were weighted by the link length and then averaged at route level by the route length. For the path overlap correction factor, an indicator variable δ_{al} is defined as $\delta_{al} =$ $1, if\ the\ link\ a\ is\ traversed\ by\ path\ l, 0\ otherwise.$ The indicator variable is summed over all the alternatives in a choice set and then weighted by the ratio of the overlapping link length to the length of the path. The weighted variable is then aggregated at route level. A detailed definition of the road attribute variables except for path overlap

correction is provided in the previous chapter, the table (Table 9) is reproduced here for the convenience of reference.

Table 9 (Reproduced from Chapter 5)

Attributes for link *a* of path *k*	Base Value	Link Level Weighted Value	Route Level Aggregated Value
Link Length	l_a	l_a	$\sum_{a\in\Gamma_k} l_a$
Annual Average Daily Traffic (AADT)	$(aadt)_a$	$(aadt)_a \times \frac{l_a}{L_k}$	$\sum_{a\in\Gamma_k} (aadt)_a \times \frac{l_a}{L_k}$
Speed Limit	s_a	$s_a \times \frac{l_a}{L_k}$	$\sum_{a\in\Gamma_k} s_a \times \frac{l_a}{L_k}$
Number of Lanes	n_a $(ranked\ categorical)$	$n_a \times \frac{l_a}{L_k}$	$\sum_{a\in\Gamma_k} n_a \times \frac{l_a}{L_k}$
Slope	$m_a = m_j - m_i$ $(m_i = elevation\ at\ start\ node,$ $m_j = elevation\ at\ end\ node)$	$m_a \times \frac{l_a}{L_k}$	$\sum_{a\in\Gamma_k} m_a \times \frac{l_a}{L_k}$
Percent of Truck	p_a	$p_a \times \frac{l_a}{L_k}$	$\sum_{a\in\Gamma_k} p_a \times \frac{l_a}{L_k}$
Presence of Bicycle Facility	δ_b $Binary(0,1)$	$\delta_b \times l_a$	$\frac{1}{L_k}\sum_{a\in\Gamma_k} \delta_b \times l_a$
Presence of Sidewalk	δ_{sw} $Binary(0,1)$	$\delta_{sw} \times l_a$	$\frac{1}{L_k}\sum_{a\in\Gamma_k} \delta_{sw} \times l_a$
Traffic Stress	$ts_a = (aadt)_a \times s_a$	$ts_a \times \frac{l_a}{L_k}$	$\sum_{a\in\Gamma_k} ts_a \times \frac{l_a}{L_k}$
Path Overlap Correction (Bovy 2008)	PSC_k	$\frac{l_a}{L_k} ln \sum_{l\in C} \delta_{al}$	$\sum_{a\in\Gamma_k} \frac{l_a}{L_k} ln \sum_{l\in C} \delta_{al}$

Multivariate Analysis

As with the deviation from the shortest route model, one trip per user was used for the route choice models resulting in a total of 445 trips to be analyzed. Of these, the choice set generation method failed to generate alternatives for 8 trips that are above 13 miles in length, further reducing the dataset to 437 trips. Although the attempt was to have at least 10 alternatives for each chosen route, the number of alternatives that were generated varied between 10 and 6, including the chosen trip. Therefore, to maintain a set

of same number of alternatives for all trips, each chosen trip was designed to have 5 alternatives. For cases where more than 5 alternatives were generated, only the first 5 were retained as the choice set generation method uses k-shortest path ideology with utility as the link cost – the further in the generation process a route is generated, the further it may be assumed to be from the optimum choice.

Among the sociodemographic variables, about 32% of the riders are in the category of comfortable but cautious and interested, but concerned, about 22% in the strong and fearless category and 46% in the category of enthused and confident which is similar to the original distribution of the rider types in the user dataset with slight oversampling of enthused and confident riders (originally 45%) and strong and fearless riders (originally 19%) and undersampling of comfortable but cautious and interested but concerned group (originally 36%). 27% of the trips are from female riders while the remaining 72% are from male riders which is similar to the distribution in the original dataset. Similarly, about 70% of the trips are from riders in the age group less than 45 years while the remaining 30% are from the older population.

Given the biased nature of the dataset and previous research suggesting that there may be significant differences in preferences between male and female riders as well as younger and older riders, it was decided to run segmented models on the groups and then a pooled model. The method followed was: (1) run separate path size logit models on subsets of data which consist only of male riders or female riders/ younger or older riders and find the best specifications for those models, (2) run pooled models on the entire data set and find the best specifications for that model, (3) for variables that are significant in one of the segmented models but not in the pooled model, or if the coefficient estimates

are significantly different (tested by t-tests), add them as an interaction term in the final pooled model. For example, if AADT is found to be significant in the female only model but not in the male only model or the pooled model, then in the next stage, a pooled model is created with an interaction term female × AADT. If that term is found to be significant in the pooled model only then it is concluded that the variable is significant for that group. Variables were added to the model sequentially and removed stepwise from the model when found insignificant. The final segmented models for male and female riders and the pooled model are presented in Table 12 while the segmented models for the age groups and their corresponding pooled model results are presented in Table 13. The t- statistic for difference in coefficients between segments was calculated as

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{var(\hat{\beta}_1) + var(\hat{\beta}_2)}}$$

To test whether the models are significantly different, χ2 test was performed as

$$-2[LL(pooled\ model) - LL(segment\ 1) - LL(segment\ 2)] \sim \chi^2_{df},$$

where the degrees of freedom (df) are equal to $K_1 + K_2 - K$ where K is the number of coefficients in the pooled model and K_s is the number of coefficients in the sth market segment model, s = 1, 2 (Koppelman and Bhat 2006).

Model Results

The model estimated only on female riders had a sample size of 119 while that with the male riders had 318. The pooled model was estimated on the total sample size of

437. Along with AADT, speed, slope, bicycle facility, number of lanes, trip distance and path overlap correction factor, interaction terms of negative road attributes (AADT, speed, slope) with bicycle facility was experimented with but was not found to be significant. AADT was found to be significant across all of the models except for the male only model and the initial pooled model for gender. Speed, however, was significant across all models as were presence of bicycle facility, number of lanes and the path size overlap correction factor. Slope was significant in the segmented model for the higher age group and retained its significance in the corresponding pooled model. For gender segmented models, age and for age segmented models, gender was included as explanatory variable but was not found to be significant and was hence not retained in the final model specification.

Between the two gender segmented models, although speed was significant in both, the coefficients were significantly different ($p = 0.002$) between the two groups. Therefore, interaction term of speed and female was added in the final pooled model. As mentioned before, AADT was significant for female riders but not for the male riders, so an interaction term between AADT and female was added to the final pooled model. Bicycle facility, number of lanes and path size correction factor were significant in all three models and were therefore retained in the final pooled model. The model fits as given by McFadden's ρ^2 (base market share model) range between 0.4 and 0.49 for all the models, which is considered relatively good for disaggregate discrete choice models of travel behavior. The odds ratio calculated shows that a unit (1000 vehicles) increase in AADT can decrease the odds of choosing a route by 55 % for females. Similarly, a unit increase (10 mph increase) in speed can decrease the odds of choosing a route by 40% for

females and by about 20% for male users. A unit increase in the percent presence of bicycle facility(for example, an increase in percent presence of bicycle facility from 0.7 to 0.8) can increase the odds of choosing that route by almost 7 times for female riders and by about 4 times for male riders. Increase in path overlap also positively affects the possibility of choosing a route which points to the fact that users generally prefer closely related paths. Significance of speed in the pooled model indicates that female riders are significantly more negatively affected by speed compared to their male counterparts, all others being constant. The $\chi 2$ test of significance for model fit is $-2[LL(final\ pooled\ model) - LL(male) - LL(female)] \sim 54.32_{11}$ which has a corresponding p –value < 0.005. Therefore, the null hypothesis that the pooled model is equivalent to the segmented models can be rejected at a significance level of 0.005. We therefore, choose the gender segmented models as better representation of the sample.

Table 12. Gender-Segmented and Pooled Multinomial Logit Route Choice Models

	Model 1		Model 2		Model 3	Model 4	
Variables	Female Only (N = 119)	Odds Ratio	Male Only (N = 318)	Odds Ratio	Initial Pooled (N = 437)	Final Pooled (N = 437)	Odds Ratio
ASC_alternative 2	**-0.864*** *(-1.91)*	0.421	**-0.934**** *(-2.89)*	0.393	**-1.1**** *(-2.66)*	**-1.78***** *(-3.4)*	0.169
ASC_alternative 3	**-0.968 *** *(-2.35)*	0.380	**-1.695*** *(-2.06)*	0.184	**-1.468.** *(-1.83)*	**-1.38**** *(-2.8)*	0.252
ASC_alternative 4	**-1.249** *(-1.46)*	0.287	**-1.668***** *(-4.15)*	0.189	**-2.52***** *(-2.94)*	**-2.12*** *(-1.95)*	0.120
ASC_alternative 5	**-1.545***** *(-3.19)*	0.213	**-1.485***** *(-3.28)*	0.227	**-2.89***** *(-3.17)*	**-3.13 ***** *(-2.89)*	0.044
ASC_chosen	**-3.291***** *(-3.23)*	0.037	**-1.36*** *(-2.56)*	0.257	**-2.24*** *(-2.462)*	**-2.76***** *(-3.22)*	0.063
Distance	**-0.72***** *(-4.37)*	0.487	**-0.4**** *(-2.68)*	0.670	**-0.56**** *(-2.79)*	**-0.62***** *(-3.18)*	0.538
AADT	**-0.8***** *(3.42)*	0.449					
Speed	**-0.5***** *(-4.02)*	0.607	-**0.14*** *(-2.33)*	0.869	**-0.3*** *(-2.57)*	**-0.187*** (2.12)	
Percent bicycle facility	**1.942***** *(8.267)*	6.973	**1.342***** *(6.27)*	3.827	**1.116***** *(5.32)*	**1.262***** *(5.72)*	3.532
Average number of lanes	**-0.401***** *(-5.21)*	0.670	**-0.47***** *(-3.205)*	0.625	**-0.42*** *(-2.19)*	**-0.4*** *(-2.47)*	0.670
Slope	**-0.41**** *(-2.76)*	0.664	**-0.127.** *(-1.93)*	0.881	**-0.29*** *(-2.21)*	**-0.32**** *(-2.85)*	0.726
AADT x female						**-0.357*** *(-2.52)*	0.811
Speed x female						**-0.28*** *(-2.12)*	
Log PS	**1.275***** *(3.27)*		**2.176***** *(3.45)*		1.973 (2.83)	**1.99**** *(2.8)*	
Model Statistics							
LL(0)	-361.22		-967.92		-1328.34	-1328.34	
LL(MS)	-321.7		-859.67		-1181.37	-1181.37	
LL(Model)	-191.55		-436.624		-668.223	-654.341	
McFadden's ρ2(base MS)	0.4		0.49		0.434	0.45	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 13. Age-Segmented and Pooled Multinomial Logit Route Choice Models

	Model 1		Model 2		Model 3	Model 4	
Variables	Age≥45 Only (N = 131)	Odds Ratio	Age <45 Only (N = 306)	Odds Ratio	Initial Pooled (N = 437)	Final Pooled (N = 437)	Odds Ratio
ASC_alternative 2	**0.45 .** *(1.86)*	1.569	-1.699 *(-0.961)*	0.183	**-1.23** *(-1.32)*	**-1.752*** *(2.28)*	0.173
ASC_alternative 3	-3.693*** (-7.36)	0.025	**-0.886**** *(-2.544)*	0.412	-1.526 (-0.724)	**-2.154.** *(-1.27)*	0.116
ASC_alternative 4	**-3.654***** *(-7.301)*	0.026	**-1.147*** *(-2.02)*	0.318	**-1.47***** *(-2.89)*	**-1.873*** *(-2.732)*	0.154
ASC_alternative 5	-0.460 *(-0.91)*	0.631	**-1.353*** *(-2.08)*	0.258	**-1.85 **** *(-2.52)*	**-1.766*** *(-2.94)*	0.171
ASC_chosen	**-1.633***** *(-2.544)*	0.195	**-1.765*** *(-2.25)*	0.171	**-2.4723*** *(-2.296)*	**-2.162*** *(-2.147)*	0.115
Distance	**-0.396***** *(-2.724)*	0.673	**-0.247*** *(-1.89)*	0.781	**-0.278*** *(-1.986)*	**-0.275*** *(-2.654)*	0.760
AADT	**-0.851***** *(-6.11)*	0.427	**-0.291**** *(-2.89)*	0.748	**-0.553**** (-2.543)	**-0.317**** *(2.37)*	
Speed	**-0.474**** *(-2.32)*	0.622	**-0.164.** *(1.72)*	0.849	**-0.175.** *(-1.69)*	**-0.18.** *(1.92)*	
Percent bicycle facility	**1.837***** *(2.763)*	6.283	**2.061*** *(2.12)*	7.858	**2.002**** *(2.76)*	**2.098***** *(2.392)*	8.150
Average number of lanes	**-0.106**** *(2.06)*	0.899	**-0.165**** *(2.35)*	0.848	**-0.139*** *(-2.23)*	**-0.16** *(-1.74)*	0.852
Slope	**-0.894***** *(-5.746)*	0.409					
AADT x age≥ 45						**-0.432**** *(-2.984)*	
Speed x age≥ 45						**-0.127**** *(-3.27)*	
Slope x age≥ 45						**-0.881***** *(-4.317)*	
Log PS	**1.456***** *(4.532)*		**1.222***** *(3.762)*		**1.692***** *(3.85)*	**1.26***** *(4.14)*	
Model Statistics							
LL(0)	-411.240		-1078.500		-1489.740	-1489.740	
LL(MS)	-343.2		-872.35		-1215.53	-1215.53	
LL(Model)	-223.19		-547.24		-863.223	-801.61	
McFadden's ρ2(base MS)	0.35		0.37		0.29	0.34	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Table 14. Rider type -Segmented and Pooled Multinomial Logit Route Choice Models

	Model 1		Model 2		Model 3		Model 4	
Variables	**Strong and fearless (N = 96)**	**Odds Ratio**	**Enthused and confident (N = 201)**	**Odds Ratio**	**Comfortable, but cautious and Interested, but concerned (N = 140)**	**Odds Ratio**	**Final Pooled (N = 437)**	**Odds Ratio**
ASC_alternative 2	-1.528 *(-1.32)*	0.217	**-1.21*** *(-1.913)*	0.298	**-1.743*** *(-2.13)*	0.175	**-1.562*** *(-2.02)*	0.210
ASC_alternative 3	**-0.723 **** *(-2.97)*	0.485	**-1.425**** *(-2.73)*	0.241	**-0.957*** *(-2.62)*	0.384	**-1.274**** *(-2.76)*	0.280
ASC_alternative 4	-1.382 *(-1.256)*	0.251	**-1.732*** *(-2.23)*	0.177	**-2.467**** *(-2.788)*	0.085	**-2.31*** *(-2.19)*	0.099
ASC_alternative 5	**-0.873*** *(-2.13)*	0.418	**-1.314**** *(-2.923)*	0.269	**-1.52**** *(-2.687)*	0.298	**-1.621 **** *(-2.56)*	0.198
ASC_chosen	**-2.661**** *(-2.72)*	0.070	**-1.522**** *(-2.62)*	0.218	**-1.923**** *(-2.834)*	0.146	**-2.28**** *(-2.78)*	0.102
Distance	**-0.37*** *(-2.25)*	0.691	**-0.41**** *(-2.68)*	0.664	**-0.48**** *(-2.79)*	0.619	**-0.422**** *(-2.98)*	0.656
AADT	**-0.54***** *(3.17)*	0.583	**-0.58***** *(3.279)*	0.560	**-0.61***** *(3.42)*	0.543	**-0.59***** *(3.21)*	0.554
Speed	**-0.23***** *(-3.682)*	0.795	**-0.258***** *(-3.34)*	0.773	**-0.31***** *(-3.77)*	0.733	**-0.264***** *(3.562)*	0.768
Percent bicycle facility	**1.146***** *(6.127)*	3.146	**1.024***** *(5.39)*	2.784	**1.231***** *(6.67)*	3.425	**1.189***** *(5.89)*	3.284
Average number of lanes	**-0.21***** *(-3.244)*	0.811	**-0.263***** *(-3.521)*	0.769	**-0.278***** *(-3.43)*	0.757	**-0.276*** *(-3.47)*	0.759
Slope	**-0.26***** *(-3.129)*	0.771	**-0.28***** *(-3.73)*	0.756	**-0.312***** *(-3.64)*	0.732	**-0.3***** *(-3.651)*	0.741
Log PS	**1.112***** *(3.02)*		**1.162***** *(3.52)*		**1.237***** *(3.76)*	0.290	**1.99***** *(3.77)*	
Model Statistics								
LL(0)	-357.404		-748.315		-521.214		-1626.933	
LL(MS)	-217.34		-452.752		-373.29		-1043.38	
LL(Model)	-171.21		-327.317		-279.52		-781.064	
McFadden's ρ2(base MS)	0.212		0.28		0.251		0.25	

*Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

For the age segmented models, except for slope, all other variables were significant in all the models. Slope is only significant for riders in the older age group. The coefficients of both AADT and speed are significantly different across the two age groups and the significance is retained in the pooled model via the interaction terms.

Thus, the final pooled model for age consists of variables AADT, speed, AADT x age group dummy, speed × age group dummy, slope × age group dummy, bicycle facility, average number of lanes and path size correction term. The ρ^2 (base MS) ranges between 0.32 and 0.39. From the odds ratios, 1000 vehicles increase in AADT decreases the odds of choosing a route by about 60% for an older rider while 10 mph increase in speed limit decreases the odds of choosing the route by 40%. Both AADT and speed affect older riders significantly more than younger riders for whom the corresponding changes are 25% and 15%. Slope also significantly deters older riders from choosing a route and by about 60% for a unit increase (1 feet) in slope.

To test the hypothesis that the pooled model is the same as the segmented models, the $\chi 2$ test was performed as $-2[LL(final\ pooled\ model) - LL(Age \geq 45) - LL(Age < 45)] \sim 31.84_{10}$ which is significant at 0.005 level indicating that the hypothesis is rejected with 99.99% confidence and the segmented model describes the sample better than the pooled model.

Segmented models were also done on rider types (Table 14). However, none of the coefficient estimates were significantly different across the groups and the $\chi 2$ test performed was significant only at the 0.9 level indicating that the segmented models and the pooled model are similarly effective in describing the sample. This result was counterintuitive as we expected route choice to be related to rider type. One possible reason may be a similar underrepresentation of females and senior riders in the rider type segments as in the population which masks the gender and age effect. In addition, since the cyclists in Atlanta do not have sufficient route choices that can enable them to be on the shortest route and yet be on bicycle facilities or on quieter streets, they are often

captive to those limited choices which is reflected in revealed preference data. In spite of their self-classification into lower confidence categories, *comfortable, but cautious* cyclists might be using the same routes as the *strong and fearless* cyclists as they might not be willing to take a detour of about 30% to access separate facilities. Since this sample did not have a substantial proportion of *interested, but concerned* riders, it is difficult to ascertain how and if their route preference varies from these other three groups.

Discussion

From the segmented models developed and discussed in the earlier section, it is evident that choice of route varies significantly among different types of cyclists – both by age and by gender. Coefficients for variables that can negatively affect route choice (e.g. AADT, speed, slope), are consistently higher for women and older riders, whether significant or not. However, since majority of cyclists in US are male and young cyclists, any data on cyclists are likely to be weighted in favor of such cyclists and their preferences. Segmented models help to separate out the preferences and choices of less represented groups.

From a policy point of view, multiple level-of-service measures and types of infrastructure are designed based on the classification of cyclists as experienced or new. Most infrastructure created for new cyclists tends to be off-road and away from the main arterials and therefore involves longer detours for any trip. Experienced bicyclists are expected to share the road comfortably with vehicular traffic with minimum or no protection. However, as this research shows, older riders and females, are deterred severely by traffic conditions and are also not willing to take longer detours. Any

infrastructure created far away from the normal commute or travel road will always fail to attract such cyclists and will remain underused. Therefore, in order to encourage riders, agencies need to plan for infrastructure that is close to the shortest routes and yet provide sufficient buffer to the cyclists from the vehicular traffic.

It should also be noted that although we performed segmentation based on rider type, there was no distinct difference between the preferences among different groups. This may be due to the biased nature of the dataset which is heavily dominated by male and younger riders. A classification by rider type does not preclude the possibility of oversampling those groups particularly because we are also oversampling the more confident riders (see the chapter on shortest route). Therefore, it may not be unexpected that preferences across rider types do not vary significantly and is dominated by the preferences of male and younger riders. In addition, as mentioned earlier, since this is a revealed preference dataset from a city without much choices for avid bicyclists, we are possibly capturing a forced choice rather than preference which may not differ significantly across the first three rider types who, although safety concerned, still identifies themselves as cyclists and would not consider mode shift easily.

Conclusion

Traditionally route choice models have considered only route attributes to be determinant of a rider's decision to choose a route. However, the segmented route choice models developed in this chapter indicate distinctly different preferences and effects of road attributes across age and gender. This result is not always evident in cyclist datasets as such data are overwhelmingly dominated by motivated cyclists who, in the US, are male and young cyclists in 70% of the cases. Therefore, developing different models for

different socio-demographic groups may be necessary to understand the actual effect of any infrastructure related decision.

One caveat of this research is that segmentation was performed deterministically, by subsetting the data into two or more premeditated groups. Often, this is fraught with preconceived ideas and philosophies on the part of the analyst. A preferred way of approaching segmentation is to let the segmentation happen probabilistically through finite mixture latent class models where class membership is not rigid or predetermined. Future research will conduct the same route choice analysis with latent class analysis and compare the segmentation with the deterministic segmentation presented in this chapter.

CHAPTER 8

CONCLUDING REMARKS AND FUTURE DIRECTION

The Cycle Atlanta project was created to provide a virtual civic engagement platform for cyclists in Atlanta where people from all walks of life could come together and participate in the planning process through data. Atlanta, not typically known as a bicyclist friendly city, has seen a growth in bicycling from 0.33% to 1.1% from 2000 to 2009 (League of American Bicyclists, 2016). However, facility creation for cyclists has not kept up with that growth rate resulting in higher friction among vehicular traffic and cyclists. In some cases, like the Ponce de Leon arterial, a buffered bike lane has been created without any lateral separation from high speed and volume traffic – it is only natural that such facilities will remain underutilized and safety concerned cyclists will refrain from using such facilities. This gap in what cyclists want and need versus what cyclists get arises because planning for cyclist infrastructure is not data informed as data on cyclists' routes and volumes are rare. The purpose of this dissertation was to close that information gap and act as a bridge between the cyclists and the Atlanta city planners by identifying patterns and trends of cyclist behaviors and preferences from the data provided by the cyclists themselves and to develop interpretable and usable models for the city planners for future use.

This dissertation is based off data collected via a crowdsourcing system and makes a contribution in that area via creating a framework for the classification of crowdsourcing systems across all disciplines and situating the current research in its context, setting a precedence for future crowdsourced transportation data collection systems. Towards cyclists and their classification, this research explored at depth (1) how

self-perception of comfort and confidence of cyclists is influenced by their age, gender, cycling frequency, experience and other factors, (2) a cyclist's propensity to choose the shortest route and how far she deviates from that route and (3) how route choice varies between different types of cyclists and if the choices are different enough to warrant separate methods of analysis.

Future Research

One of the factors that have not been accounted for in this research is the effect of the different types of bicycle facilities on route choice. Previous research and literature differ widely in opinions on this aspect although the majority agree that bike lanes on shared facilities without physical barriers between vehicle lanes and bicycle lanes do not make much impact on cyclists' perception of safety. In the future, the route choice study presented here should be extended to include different types of bicycle facilities, so that the differential effectiveness of the facilities can be identified. In addition, this study did not distinguish between presence of turn lanes and did not account for signalized intersections or the wait time at such intersections, both of which can potentially affect link level choice decisions. Future route choice models should address these issues to derive a fuller model specification.

This research developed a map matching algorithm using network information to reduce the computational burden of the nearest neighbor search based map matching algorithms. However, the algorithm needs to be validated for its computational efficiency and snapping accuracy and compared with the existing methodologies to establish its uniqueness. Similar validation is also required for the choice set generation algorithm which will be carried out in the near future.

A valuable extension of this research will be to model route choice decisions as link level decisions instead of route level decisions using a Bayesian framework. First, except for commuter cyclists, and particularly for new cyclists, route level information is not always complete and decisions are made at a link level and in real time rather than at route level. Modelling the route choice decision as a Bayesian process will also give us the ability to compare the two methods on their power of prediction and future research can be informed accordingly. In addition, if it is found that cyclists do not actually strictly adhere to a premeditated route, it will provide strong reasons for including route suggestions and/or real time information updates n route recording applications like Cycle Atlanta which may help in user retention. Second, when modelled at the route level, the significant role that some variables like slope, play at the link level decision gets masked. Therefore, to identify the actual effect that link attributes have on route choice of cyclists, a link level decision making process might be a more reasonable approach.

Two immediate future research ideas, however, have already been identified and are in their preliminary stages. The first is replacing the deterministic segmentation of cyclists by a probabilistic latent classification for the route choice models and the second is generating differential choice sets for different segments of users with random coefficients calibrated to represent taste variations across the segments.

Conclusion

The overarching theme of this dissertation is that all cyclists are not created equal and their behaviors and attitudes are not singularly defined by either infrastructure or by their comfort and confidence level or by their experience and cycling frequency. How a cyclist will respond to infrastructure and how far they will be willing to travel to access that infrastructure is a complex decision process, only a part of which can be realistically estimated through collecting stated preference data on cyclists. Revealed preference data collection, however, is resource intensive and agencies shy away from it for small and marginal users of the transportation system like cyclists. Crowdsourced data collection systems provide the happy solution where agencies do not have to invest extensive resources and yet can get revealed preference data from the users directly. The system is not without its caveats though – only motivated and interested people participate in crowdsourcing systems, thereby introducing self-selection bias in the data. Additionally, data qualities are not always up to a standard and dedicated data processing efforts are required to filter useful information from the collected data. However, the civic participation mobilized by crowdsourcing systems considerably outweighs its disadvantages which can be handled by preplanning and designing specifically to address such issues.

The results from this research provide valuable insight for future planning and policy decisions. First, female and senior cyclists are found to be in general low confidence, low comfort riders and they significantly differ in their route choice and infrastructure preference from their more confident counterparts. Therefore, building infrastructure based on cycling experience alone, is not sufficient to address the needs of

all cyclists. Second, the assumption that with more riding experience cyclists become confident enough to share the street with vehicular traffic, is not without its caveats. Although cyclists with more riding experience tend to see themselves as more confident riders, preference for separate infrastructure pervades all rider types, as does the negative influence of high speed and high volume traffic. Third, cyclists are generally found to shy away from longer trips and hence, when faced with the trade-off between a significant detour and safety concerns, they may not make the trip itself. Therefore, having a connected network close to the shortest distance path is important in encouraging newer and low confidence bicyclists. This research provides a model that can be used to estimate acceptable deviation from any route based on road attributes and the cyclist characteristics. It is also noted that while there was significant differences in route preference among male and female riders, and among younger and older riders, there is no such difference among rider types, which are assumed to reflect cyclists' confidence and comfort with different infrastructure. However, since this is revealed preference data from a city with less choices for bicyclists, it is quite possible that what we are modelling is a forced choice rather than preference. Low presence of interested but concerned riders, i.e., riders with substantially low confidence in the dataset attests to the fact that present state of infrastructure in Atlanta does not provide enough safe choices for such riders and what we are capturing are essentially riders who somewhat identify themselves as cyclists, even with safety concerns and would not consider mode shift easily. This finding has a huge policy implication in that if we actually want to promote cycling as a mode of transportation and see substantial growth, we need to build enough infrastructure such that people have sufficient and varied choices that make them feel safe to take up

bicycling. Coupled with the findings from the shortest route model, this implies finding alternative routes to main arterials that are within 5-10% longer than the shortest route as seen in other bicycling cities like Portland and San Francisco, to build separate bicycling infrastructure, initiating traffic calming measures along all major arterials and maintaining a connected network of bicycling facilities.

In the future, the models developed in this dissertation will be extended to include more link level attributes for a fuller model specification along with identifying probabilistic latent classification of cyclists for the purpose of route choice. Modelling route choice as a link level decision rather than a path level decision is also a potential future exploration which is expected to give a better understanding of possible deviation of the cyclists from the shortest path.

REFERENCES

1. Innes, J. E. Information in Communicative Planning, *Journal of the American Planning Association*, Vol 64, No.1, pp.52-63,1998.
2. Burby, R.J. Making Plans that Matter, *Journal of the American Planning Association*, Vol.69, No.1, pp.33-49, 2003.
3. Slotterback, C.S. Public Involvement In Transportation Project Planning and Design, *Journal of Architectural and Planning Research*, Vol.27, No.2, pp.144, 2010.
4. Insua. R. D., Kersten E.G., Rios J. and Grima C. Towards decision support for participatory democracy,*ISeB*, Vol. 6, pp.161–191, 2008.
5. Hague, C., Kirk, K., Higgins, M., Prior, A., Jenkins, P., Smith, H. and Grimes, W. Participatory planning for sustainable communities. 2003.
6. Rabinowitz, P. Participatory Approaches to Planning Community Interventions, *http://ctb.ku.edu/en/tablecontents/sub_section_main_1143.aspx*, 2013. Accessed May 2013.
7. Skocpol T., and Fiorina, M. (eds.) Civic Engagement in American Democracy. Brookings Institution Press,1999.
8. Galston,W. A. Civic education and political participation. *PS: Political Science & Politics*, Vol.37, pp. 253–266, 2004.
9. Pew Research Center for the People and the Press. Cable and Internet loom large in fragmented political news universe. *http://www.people-press.org/2004/01/11/cable-and-internet-loom-large-in-fragmented-political-newsuniverse/*, 2004. Accessed May 2013.
10. Wagner, J. Measuring the Performance of Public Engagement in Transportation Planning: Three Best Principles. In *TRB 2013 Annual Meeting*, Vol. 954, 2012.
11. Howe, J. The Rise of Crowdsourcing. *Wired, Conde Nast Digital*, 14, 6, Jun, 2006.

12. Doan, A,. Ramakrishnan, R., Halevy, A.Y. Crowdsouring Systems on the World-Wide Web, *Communications of the ACM*. Vol. 54, No. 4, pp. 86-96, 2011.
13. Saxton, G. D., Oh, O.,and Kishore, R. Rules of Crowdsourcing: Models, Issues, and Systems of Control, *http://www.acsu.buffalo.edu/~rkishore/papers/Saxton-et-al-Crowdsourcing-ISM-Forthcoming.pdf* . Accessed May 2013.
14. Goodchild, M. Assertion and authority: the science of user-generated geographic content. *Earth*, 2008, pp. 1-18.
15. Kuznetsov,S. and Paulos, E. Participatory Sensing in Public Spaces: Activating Urban Surfaces with Sensor Probes, *ACM Designing Interactive Systems (DIS)* 2010.
16. Kuznetsov, S., Davis, G. N., Cheung, J. C. and Paulos, E. Ceci N'est Pas Une Pipe Bombe: Authoring Urban Landscapes with Air Quality Sensors, *ACM SIGCHI* 2011.
17. Hood,J. , Sall, E., Charlton, B. A GPS-based bicycle route choice model for SanFrancisco, California. *Transportation Letters: The International Journal of Transportation Research*, Vol.3, Jan. 2011, pp. 63-75
18. Kitchin, R. and Dodge, M. Rethinking Maps, In *The Map Reader: Theories of Mapping Practice and Cartographic Representation (eds M. Dodge, R. Kitchin and C. Perkins)*, John Wiley & Sons, Ltd, Chichester, UK.
19. Steinfield, A., Zimmerman, J. and Tomasic, A. Bringing Customers Back into Transportation: Citizen-Driven Transit Service Innovation via Social Computing, In *Best Practices for Transportation Agency Use of Social Media*, (*eds. S. Bregman and K. Watkins)*, CRC Press, 2013
20. Erickson, T. Geocentric Crowdsourcing and Smarter Cities: Enabling Urban Intelligence in Cities and Regions. *A position paper for the 1st International workshop on ubiquitous crowdsourcing. UbiComp'10*, Copenhagen, Denmark, 2010.

21. Priedhorsky, R., Jordan, B., and Terveen, L. How a Personalized Geowiki Can Help Bicyclists Share Information More Effectively. *Proc. 2007 Int'l Symposium on Wikis, ACM*, 2007, pp. 93-98.
22. Heipke, C. Crowdsourcing Geospatial Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 65, No. 6, pp. 550-557, 2010.
23. Brabham, D. C. Crowdsourcing the Public Participation Process for Planning Projects. *Planning Theory*, Vol. 8, No. 3, Jul. 2009, pp. 242-262.
24. SeeClickFix. *http://seeclickfix.com/*, 2013. Accessed May 2013.
25. Ferris, B., Watkins, K., and Borning, A. OneBusAway: Results from Providing Real-Time Arrival Information for Public Transit. *Proceedings of CHI 2010*. Atlanta, GA, USA, April 10 – 15, 2010.
26. Zimmerman, J., Tomasic, A., Garrod, C., Yoo, D., Hiruncharoenvate, C., Aziz, R., Thiruvengadam, N.R., Huang, Y. and Steinfeld, A. Field trial of Tiramisu: crowd-sourcing bus arrival times to spur codesign. *In Proceedings of CHI '11*. *ACM*, New York, NY, USA, 1677-1686, 2011.
27. Cycle Atlanta. *http://www.cycleatlanta.org*, 2013. Accessed May 2013.
28. Dill, J. and McNeil, N. Four Types of Cyclists? Examining the Typology to Better Understand Bicycling Behavior and Potential, *92nd Annual Meeting of the Transportation Research Board*, 2013.
29. Masli, M. Crowdsourcing Maps, *Computer*, Vol. 44, No. 11, 2011, pp. 90-93.
30. Cyclopath. *http://cyclopath.org*, 2011. Accessed May 2013.
31. Wiggins, A. & Crowston, K. From conservation to crowdsourcing: A typology of citizen science,In *Proceedings of the Forty-fourth Hawai'i International Conference on System Science (HICSS- 44)*, Koloa, HI, 4–7 January.
32. Gooze, A.I. Real-time Transit Information Accuracy: Impacts and Proposed Solutions, M.S Thesis, Georgia Tech, *http://hdl.handle.net/1853/47638*, 2013. Accessed May 2013.

33. Gooze, A, Watkins, K., Borning, A. Benefits of Real-Time Information and the Impacts of Data Accuracy on the Rider Experience. To appear in *Transportation Research Record: Journal of the Transportation Research Board*, 2013.

34. Windmiller, S., Hennessy, T. and Watkins, K. "Communication Technology Usage and the Rider Experience: A Case Study of St. Louis Metro", submitted to Transportation Research Board Annual Meeting, 2014.

35. AASHTO. (2012) Guide for the Development of Bicycle Facilities.

36. Adam, N.R., and Wortman, J.C. (1989) Security-control methods for statistical databases: A comparative study. ACM Computing Surveys, 21:515-556.

37. Ardagna C.A., Cremonini M., Damiani E., Vimercati S.C, Samarati P. (2007) Location Privacy Protection Through Obfuscation-based Techniques, In, Data and Applications Security XXIV 4602 Lecture Notes in Computer Science E Barker, Steve.E Ahn, Gail-Joon.

38. Aultman-Hall, et al (1997) Analysis of Bicycle Commuter Routes Using Geographic Information Systems. Transportation Research Record No. 1578, Transportation Research Board.

39. Bayardo, Jr. R. J. and Agrawal, R. (2005) Data Privacy through Optimal k-Anonymization. In ICDE, 2005.

40. Bekhor, S., Ben-Akiva, M.E., Ramming, S., (2006). Evaluation of choice set generation algorithms for route choice models. Annals of Operations Research, 144(1), 235-247.

41. Bekhor, S., Prato, C.G., (2006). Effects of choice set composition in route choice modeling. Proceedings of the 11th International Conference on Travel Behavior Research, Kyoto, Japan.

42. Ben-Akiva, M. & Bierlaire, M. (1999), "Discrete choice methods and their applications to short term travel decisions," in R. Hall, ed., *Handbook of Transportation Science*, Kluwer Academic Publishers, Norwell, MA, chapter 2,

43. Ben-Akiva, M.E., Bergman, M.J., Daly, A.J., Ramaswamy, R., (1984). Modeling inter-urban route choice behaviour. In: Volmuller, J., Hamerslag, R. (Eds.), Proceedings of the 9th International Symposium on Transportation and Traffic Theory. VNU Science Press, Utrecht, The Netherlands, 299-330.
44. Ben-Akiva, M.E., Bolduc, D., (1996). Multinomial probit with a logit kernel and a general parametric specification of the covariance structure. Working Paper, Massachusetts Institute of Technology, Cambridge, USA.
45. Beresford, A. R. and Stajano, F. (2003) Location Privacy in Pervasive Computing. IEEE Pervasive Computing, 2(1):46–55, 2003.

    Mokbel, M. F. (2006) Towards Privacy-Aware Location-Based Database Servers. In Proceedings of the ICDE International Workshop on Privacy Data Management, PDM, 2006.
46. Bierlaire, M., Frejinger, E., (2005). Route choice models with subpath components. Proceedings of the 5th Swiss Transport Research Conference, Ascona, Switzerland.
47. Bierlaire, M., Frejinger, E., (2008). Route choice modeling with network-free data. Transportation Research Part C, 16(2), 187-198.
48. Bliemer, M.C.J., Bovy, P.H.L., (2008). Impact of route choice set on route choice probabilities. Transportation Research Record, 2076, 10-19.
49. Bovy, P.H.L., (2009). On modelling route choice sets in transportation networks: a synthesis. Transport Reviews, 29(1), 43-68.
50. Bovy, P.H.L., Bekhor, S., Prato, C.G., (2008). The factor of revised path size: an alternative derivation. Transportation Research Record, 2076, 132-140.
51. Bovy, P.H.L., Bekhor, S., Prato, C.G., (2009). Route sampling correction for stochastic route choice set generation. Proceedings of the 88th Annual Meeting of the Transportation Research Board, Washington, D.C.

52. Bovy, P.H.L., Fiorenzo-Catalano, S., (2007). Stochastic route choice set generation: behavioral and probabilistic foundations. Transportmetrica, 3(3), 173-189.
53. Broach, J, et al (2010) Bicycle route choice model developed using revealed preference GPS data. Transportation Research Board Annual Meeting Compendium.
54. Buehler, R. and Pucher. J. (2011) Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes, http://policy.rutgers.edu/faculty/pucher/bikepaths.pdf, accessed August 2013.
55. Cascetta, E., (2001). Transportation Systems Engineering: Theory and Methods. Kluwer Academic Publishers, Dordrecht, The Netherlands.
56. Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A., (1996). A modified logit route choice model overcoming path overlapping problems: specification and some calibration results for interurban networks. In: Lesort, J.B. (Ed.), Proceedings of the Thirteenth International Symposium on Transportation and Traffic Theory. Pergamon, Lyon, France, 697-711.
57. Charlton, B, et al (2010) CycleTracks – a Bicycle Route Choice Data Collection Application for GPS-Enabled Smart Phones. Innovations in Travel Modeling paper compendium, Transportation Research Board.
58. Chu, C., (1989). A paired combinatorial logit model for travel demand analysis. Proceedings of the 5th World Conference on Transportation Research, Ventura, USA, 295-309.
59. Daganzo, C.F., Sheffi, Y., (1977). On stochastic models of traffic assignment. Transportation Science, 11, 253-274.
60. De la Barra, T., Perez, B., Anez, J., (1993). Multidimensional path search and assignment. Proceedings of the 21st PTRC Summer Annual Meeting, Manchester, England, 307-319.
61. Dial, R.B., (1971). A probabilistic multipath traffic assignment model which obviates path enumeration. Transportation Research, 5(2), 83-111.

62. Dijkstra, E.W., (1959). A note on two problems in connection with graphs. Numerical Mathematics, 1, 269-271.
63. Dill, J, et al (2008) Understanding and Measuring Bicycling Behavior: A Focus on Travel Time and Route Choice. Oregon Transportation Research and Education Consortium, final report OTREC-RR-08-03.
64. Dill, J. (2004) Measuring Network Connectivity for Bicycling and Walking.
65. Dill, J. and Carr, T. (2003) Bicycle Commuting and Facilities in Major U.S.. Cities: If You Build Them, Commuters Will Use Them, Transportation Research Records, Volume 1828, pg 116-123.
66. Dill, J. and McNeil, N. (2012) Four types of cyclists? Examining the typology to better understand bicycling behavior and potential., in 92nd Annual Meeting of the Transportation Research Board.
67. Dougherty, M., (1995). A review of neural networks applied to transport. Transportation Research Part C, 3(4), 247-260.
68. Eppstein, D., (1998). Finding the *K* shortest paths. Journal of the Society for Industrial and Applied Mathematics, 28(2), 652-673.
69. Fiorenzo-Catalano, S., Van der Zij, N.J., (2001). A forecasting model for inland navigation based on route enumeration. Proceedings of the European Transport Conference, PTRC Education and Research Services Ltd., London, 1-11.
70. Frejinger, E., (2007). Random sampling of alternatives in a route choice context. Proceedings of the European Transport Conference, Leeuwenhorst, The Netherlands.
71. Frejinger, E., Bierlaire, M., (2007). Capturing correlation with subnetworks in route choice models. Transportation Research Part B, 41(3), 363-378.
72. Frejinger, E., Bierlaire, M., (2007). Capturing correlation with subnetworks in route choice models. Transportation Research Part B, 41(3), 363-378.

73. Frejinger, E., Bierlaire, M., Ben-Akiva, M.E., (2009). Sampling of alternatives for route choice modeling. Transportation Research Part B,
74. Friedrich, M., Hofsäß, I., Wekeck, S., (2001). Timetable-based transit assignment using branch & bound. Transportation Research Record, 1752, 100-107.
75. Gliebe, J.P., Koelman, F., Ziliaskopoulos, A., (1999). Route choice using a paired combinatorial logit model. Proceedings of the 78th Annual Meeting of the Transportation Research Board, Washington, D.C.
76. Hoh B., Gruteser M., Herring R., Ban, J., Work, D. Herrera, J-C, Bayen, A, M., Annavaram, M., Jacobson, Q. (2008) Virtual Trip Lines for Distributed Privacy-Preserving Traffic Monitoring, *MobiSys'08*
77. Hood, J. et al. (2011) A GPS-based bicycle route choice model for San Francisco, California. Transportation Letters: the International Journal of Transportation Research 3: 63-75.
78. Horowitz, J.L., Louviere, J.J., (1995) What is the role of consideration sets in choice modeling? International Journal of Research in Marketing, 12, 39-54.
79. Hudson, J, et al (2012) Using Smartphones to Collect Bicycle Travel Data in Texas. University Transportation Center for Mobility, Texas Transportation Institute, Project report UTCM 11-35-69.
80. Hunt, D.T., Kornhauser, A.L., (1997). Assigning traffic over essentially-least-cost paths. Transportation Research Record, 1556, 1-7.
81. Jan, O., Horowitz, A., Peng, Z., (2000). Using GPS data to understand variations in path choice. Transportation Research Record 1706, 145-151.
82. Koelman, F., Wen, C., (1998). Alternative nested logit models: structure, properties and estimation. Transportation Research Part B, 32(5), 289-298.

83. Krizek K. (2007) Two Approaches to Valuing Some of Bicycle Facilities' Presumed Benefits: Propose a session for the 2007 National Planning Conference in the City of Brotherly Love. Journal of the American Planning Association. 72(3): 309-320.
84. Kuby, M., Zhongyi, X., Xiaodong, X., (1997). A minimax method for finding the k-best differentiated paths. Geographical Analysis, 29(4), 298-313.
85. Lam, T.C., Small, K., (2001). The value of time and reliability: measurement from a value pricing experiment. Transportation Research Part E, 37(2-3), 231-251.
86. League of American Bicyclists. Bicycle Commuting Data, http://www.bikeleague.org/
87. Liu, S., Araujo, M., Brunskill, E., Rossetti, R., Barros, J., and Krishnan, R. (2013. Understanding sequential decisions via inverse reinforcement learning. In Mobile Data Management (MDM), IEEE 14th International Conference on, volume 1, pages 177–186. IEEE, 2013.
88. Lombard, K., Church, R.L., (1993). The gateway shortest path problem: generating alternative routes for a corridor location problem. Geographical Systems, 1, 25-45.
89. M. Gruteser and D. Grunwald. (2003) Anonymous Usage of Location-Based Services Through Spatial and Temporal Cloaking. In MobiSys, 2003.
Chow, C-Y, Mokbel, M.F. and Liu, X. (2006) A Peer-to-Peer Spatial Cloaking Algorithm for Anonymous Location-Based Services, ACM-GIS'06
90. Ma. C.Y.T, Yau, D.K.Y, Yip N. K., Rao, N S V (2010) Privacy Vulnerability of Published Anonymous Mobility Traces, MobiCom 2010.
91. McFadden, D., Train, T., (2000). Mixed MNL models for discrete response. Journal of Applied Econometrics, 15(5), 447-470.
92. Menghini, G. et al (2010) Route choice of cyclists in Zurich. Transportation Research Part A 44 754–765.
93. Nielsen, O.A., (2004). Behavioural responses to pricing schemes: description of the Danish AKTA experiment. Journal of Intelligent Transportation Systems, 8(4), 233-251.

94. Nielsen, O.A., Daly, A., Frederiksen, R.D., (2002). A stochastic multi-class road assignment model with distributed time and cost coefficients. Networks and Spatial Economics, 2, 327-346.
95. Opportunities for collaboration on transportation and public health research. Transportation Research Part A 28(4): 249-268.
96. Park, D., Rilett, L.R., (1997). Identifying multiple and reasonable paths in transportation networks: a heuristic approach. Transportation Research Record 1607, 31-37.
97. Poznanski, A. J. (2013), Analyzing demographic and geographic characteristics of "Cycle Atlanta" smartphone application users. MS thesis, Georgia Tech.
98. Prashker, J.N., Bekhor, S., (1998). Investigation of stochastic network loading procedures. Transportation Research Record, 1645, 94-102.
99. Prashker, J.N., Bekhor, S., (2000). Congestion, stochastic, and similarity effects in stochastic user equilibrium models. Transportation Research Record, 1733, 80-87.
100. Prashker, J.N., Bekhor, S., (2004). Route choice models used in the stochastic user equilibrium problem: a review. Transport Reviews, 24(4), 437-463.
101. Prato, C. G. (2009) Route choice modeling: past, present and future research directions, *Journal of Choice Modelling*, 2(1), pp. 65-100
102. Prato, C.G., (2005). Latent factors and route choice behaviour. Ph.D. Thesis, Turin Polytechnic, Italy.
103. Prato, C.G., Bekhor, S., (2006). Applying branch & bound technique to route choice set generation. Transportation Research Record, 1985, 19-28.
104. Prato, C.G., Bekhor, S., (2007). Modeling route choice behavior: how relevant is the choice set composition? Transportation Research Record, 2003, 64-73.
105. Pucher, J. and R. Buehler. (2007) Making Cycling Irresistible: Lessons from the Netherlands, Denmark and Germany. Transport Reviews: A Transnational Transdisciplinary Journal. 28(4): 495-528.

106. Ramming, S.,(2002). Network knowledge and route choice. Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, USA.

107. Ruphail, N.M., Ranjithan, S.R., ElDessouki, W., Smith, T., Brill, E.D., (1995). A decision suort system for dynamic pre-trip route planning. Applications of advanced technologies. In: Transportation Engineering: Proceedings of The Fourth International Conference, 325-329.

108. Sallis, JF, Frank LD, Saelens BE, Kraft MK (2004) Active transportation and physical activity:

109. Samarati P. and Sweeney, L. (1998). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, Computer Science Laboratory, SRI International,

110. Schussler, N. & Axhausen, K. (2009), "Processing raw data from global positioning systems without additional information," Transportation Research Record: Journal of the Transportation Research Board 2105, 28–36.

111. Scott, K., Pabon-Jimenez, G., Bernstein, D., (1997). Finding alternatives to the best path. Proceedings of the 76th Annual Meeting of the Transportation Research Board, Washington, D.C.

112. Sener, I., Eluru, N. & Bhat, C. (2009), "An analysis of bicycle route choice preferences in Texas, US, *Transportation* 36, 511–539. Transportation Research Board Annual Meeting Compendium

113. Sheffi, Y., (1985). Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods. Prentice-Hall, Englewood Cliffs, USA.

114. Sheffi, Y., Powell, W.B., 1982. An algorithm for the equilibrium assignment problem with random link times. Networks, 12, 191-207.

115. Swait, J., (2001). Choice set generation within the generalized extreme value family of discrete choice models. Transportation Research Part B, 35(7), 643-666.

116. Swait, J., Ben-Akiva, M.E., (1987). Incorporating random constraints in discrete models of choice set generation. Transportation Research Part B, 21(2), 91-102.

117. The City of Atlanta, Department of Planning and Community Development Department (2011) CycleAtlanta:Phase1.0 http://documents.atlantaregional.com/lci/2012Applications/Innov_Atlanta_CycleAtlanta.pdf, accessed November 2013.

118. Van der Zij, N.J., Fiorenzo-Catalano, S., (2005). Path enumeration by finding the constrained K-shortest paths. Transportation Research Part B, 39(6), 545-563.

119. Vovsha, P., (1997). The cross-nested logit model: application to mode choice in the Tel Aviv metropolitan area. Transportation Research Record, 1607, 13-20.

120. Wen, C., Koelman, F., (2001). The generalized nested logit model. Transportation Research Part B, 35(7), 627-641.

121. Zhou, Z., Chen, A., (2003). Stochastic user equilibrium problem: a comparison between length-based and congestion-based C-Logit models. In: Loo, B.P.Y., Lam, S.W.K. (Eds.), Proceedings of the 8th Hong Kong Society of Transportation Studies Conference: Transportation and Logistics. Hong Kong, China, 244-253.

122. Ziebart, B. D.,Maas, A. L,, Bagnell, J.L and Dey, A. K., (2008).Maximum entropy inverse reinforcement learning. In Artificial Intelligence, AAAI 23th Conference on, pages 1433–1438,.

123. R. Buehler and J. Pucher, “Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes,” Transportation, vol. 39, no. 2, pp. 409–432, Mar. 2012.

124. J. Dill and J. Gliebe, “Understanding and measuring bicycling behavior: A focus on travel time and route choice,” 2008.

125. M. Winters, K. Teschke, M. Grant, E. M. Setton, and M. Brauer, “How Far Out of the Way Will We Travel?: Built Environment Influences on Route Selection for Bicycle and Car Travel,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2190, no. -1, pp. 1–10, Dec. 2010.

126. P. J. Krenn, P. Oja, and S. Titze, “Route choices of transport bicyclists: a comparison of actually used and shortest routes,” Int. J. Behav. Nutr. Phys. Act., vol. 11, no. 1, p. 31, 2014.

127. J. Broach, J. Dill, and J. Gliebe, “Where do cyclists ride? A route choice model developed with revealed preference GPS data,” Transp. Res. Part Policy Pract., vol. 46, no. 10, pp. 1730–1740, Dec. 2012.

128. J. Hood, E. Sall, and B. Charlton, “A GPS-based bicycle route choice model for San Francisco, California,” Transp. Lett. Int. J. Transp. Res., vol. 3, no. 1, pp. 63–75, Jan. 2011.

129. F. Godefroy and C. Morency, “Estimating Latent Cycling Trips in Montreal, Canada,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2314, no. -1, pp. 120–128, Dec. 2012.

130. J. Parkin, M. Wardman, and M. Page, “Estimation of the determinants of bicycle mode share for the journey to work using census data,” Transportation, vol. 35, no. 1, pp. 93–109, Nov. 2007.

131. M. Nankervis, “The effect of weather and climate on bicycle commuting,” Transp. Res. Part Policy Pract., vol. 33, no. 6, pp. 417–431, 1999.

132. C. Brandenburg, A. Matzarakis, and A. Arnberger, “The effects of weather on frequencies of use by commuting and recreation bicyclists,” Adv. Tour. Climatol., vol. 12, pp. 189–197, 2004.

133. L. F. Miranda-Moreno and T. Nosal, “Weather or Not to Cycle: Temporal Trends and Impact of Weather on Cycling in an Urban Environment,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2247, no. -1, pp. 42–52, Dec. 2011.

134. J. Pucher, C. Komanoff, and P. Schimek, “Bicycling renaissance in North America? Recent trends and alternative policies to promote bicycling,” Transp. Res. Part Policy Pract., vol. 33, no. No. 7/8, pp. 625–654, 1999.

135. R. Cervero and M. Duncan, “Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area,” Am. J. Public Health, vol. 93, no. 9, pp. 1478–1483, 2003.

136. M. Winters, M. Brauer, E. M. Setton, and K. Teschke, “Built Environment Influences on Healthy Transportation Choices: Bicycling versus Driving,” J. Urban Health, vol. 87, no. 6, pp. 969–993, Dec. 2010.

137. J. Dill and T. Carr, “Bicycle commuting and facilities in major US cities: if you build them, commuters will use them,” Transp. Res. Rec. J. Transp. Res. Board, vol. 1828, no. 1, pp. 116–123, 2003.

138. M. J. Smart, K. M. Ralph, B. D. Taylor, C. Turley, and A. E. Brown, “Honey, Can You Pick-Up Groceries on Your Way Home? Analyzing activities and travel among students and in non-traditional households,” UCTC-FR-2014-07, 2014.

139. J. Pucher, J. Dill, and S. Handy, “Infrastructure, programs, and policies to increase bicycling: An international review,” Prev. Med., vol. 50, pp. S106–S125, Jan. 2010.

140. M. Winters, G. Davidson, D. Kao, and K. Teschke, “Motivators and deterrents of bicycling: comparing influences on decisions to ride,” Transportation, vol. 38, no. 1, pp. 153–168, Jan. 2011.

141. A. Misra, K. Watkins, and C. A. Le Dantec, “Socio-demographic Influence on Cyclists’ Self Classification by Rider Type,” presented at the Transportation Research Board 94th Annual Meeting, 2014.

142. M. S. Urban, C. D. Porter, K. E. Proussalgoglou, R. Calix, and C. Chu, “Modeling the Impacts of Bicycle Facilities on Commute and Recreational Bicycling in Los Angeles County,” in Transportation Research Board 93rd Annual Meeting, Washington, D.C., 2014, vol. No. 14–3904.

143. J. Dill and N. McNeil, “Four Types of Cyclists?: Examination of Typology for Better Understanding of Bicycling Behavior and Potential,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2387, no. -1, pp. 129–138, Dec. 2013.

144. HCM 2010 Highway Capacity Manual, vol. 3, 4 vols. Transportation Research Board, 2010.

145. D. L. Harkey, D. W. Reinfurt, and M. Knuiman, “Development of the bicycle compatibility index,” Transp. Res. Rec. J. Transp. Res. Board, vol. 1636, no. 1, pp. 13–20, 1998.
146. San Francisco Department of Public Health Environmental Health Section, “Bicycle Environmental Quality Index Data Collection Manual,” n.d.
147. M. C. Mekuria, P. G. Furth, and H. Nixon, “Low-stress bicycling and network connectivity,” Mineta Transportation Institute, 11-19, 2012.
148. HCM, HCM 2010 Highway Capacity Manual, vol. 1, 4 vols. Transportation Research Board, 2010.
149. J. Parks, A. Tanaka, P. Ryus, C. M. Monsere, N. McNeil, and M. Goodno, “Assessment of Three Alternative Bicycle Infrastructure Quality-of-Service Metrics,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2387, no. -1, pp. 56–65, Dec. 2013.
150. M. Winters, “Improving Public Health Through Active Transportation: Understanding the Influence of the Built Environment on Decisions to Travel by Bicycle,” The University of British Columbia, 2011.
151. American Association of State Highway and Transportation Officials, “Guide for the Development of Bicycle Facilties,” AASHTO, Washington, D.C., 1999.
152. AASHTO, Guide for the Development of Bicycle Facilities, 4th Edition, 4th ed. 2012.
153. J. Dill and K. Voros, “Factors affecting bicycling demand: Initial survey findings from the Portland region,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2031, no. 1, pp. 9–17, 2007.
154. R. L. Sanders, “Dissecting perceived traffic risk as a barrier to adult bicycling,” in Proceedings of the 92nd Annual Meeting of the Transportation Research Board, Washington, D.C., 2013.
155. F. Ahmed, G. Rose, and C. Jakob, “Commuter Cyclist Travel Behavior: Examination of the Impact of Changes in Weather,” Transp. Res. Rec. J. Transp. Res. Board, vol. 2387, no. -1, pp. 76–82, Dec. 2013.
156. R. Geller, “Four Types of Cyclists.” Portland Office of Transportation, 2006.

157. U.S.. Department of Transportation Federal Highway Administration, “Highway Functional Classification Concepts, Criteria and Procedures, 2013 Edition,” USDOT FHWA, 2013.

158. C. Monsere, J. Dill, N. McNeil, K. Clifton, N. Foster, T. Goddard, M. Berkow, J. Gilpin, K. Voros, D. van Hengel, and others, “Lessons from the Green Lanes: Evaluating Protected Bike Lanes in the US,” 2014.

159. Atlanta BeltLine, “Atlanta BeltLine Overview.” 2015.

160. Greene, W. H. (2005). Censored data and truncated distributions. *Available at SSRN 825845*.

161. Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications*. Cambridge university press.

162. Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables* (Vol. 7). SAGE.

163. Amemiya, T. (1984). Tobit models: A survey. *Journal of econometrics*, *24*(1-2), 3-61.

164. Tobin, J. (1958). Estimation of Relationships for Limited Dependent Variables. *Econometrica, 26*(1), 24-36. doi:1. Retrieved from http://www.jstor.org/stable/1907382 doi:1

165. Mills, J. (1926). Table of the Ratio: Area to Bounding Ordinate, for Any Portion of Normal Curve. *Biometrika, 18*(3/4), 395-400. doi:1. Retrieved from http://www.jstor.org/stable/2331957 doi:1

166. Casello J. M., Nour A., Rewa K. C., Hill J. (2011) An analysis of stated preference and GPS data for bicycle travel forecasting. 90th Annual Meeting of Transportation Research Board

167. El-Geneidy A., Krizek K. J., Iacono M. (2007) Predicting bicycle travel speeds along different facilities using GPS data: a proof of concept model. 86th Annual Meeting of Transportation Research Board.

168. Harvey F., Krizek K.J., Collins R. (2008) Using GPS Data to Assess Bicycle Commuter Route Choice. 87th Annual Meeting of Transportation Research Board

169. Segadilha, A. B. P., Sanches, S. P. (2014a). Identification of Factors that Influence Cyclistś Route Choice, Procedia - Social and Behavioral Sciences, Volume 160, pp. 372-380

170. Segadilha, A. B. P., Sanches, S. P. (2014b). Analysis of Bicycle Commuter Routes Using GPSs and GIS, Procedia - Social and Behavioral Sciences, Volume 162, pp.198-207

171. Aultman-Hall, et al (1997) Analysis of Bicycle Commuter Routes Using Geographic Information Systems. Transportation Research Record No. 1578, Transportation Research Board.

172. Shafizadeh, K., & Niemeier, D. (1997). Bicycle journey-to-work: travel behavior characteristics and spatial attributes. *Transportation Research Record: Journal of the Transportation Research Board*, (1578), 84-90.

173. Howard, C., & Burns, E. (2001). Cycling to work in Phoenix: route choice, travel behavior, and commuter characteristics. *Transportation Research Record: Journal of the Transportation Research Board*, (1773), 39-46.

174. Krenn PK, Titze S, Oja P, Jones A, Ogilvie D: Use of global positioning systems (GPS) to study physical activity and the environment: a systematic review. Am J Prev Med 2011, 41(5):508–515.

175. Hunt, J. D., & Abraham, J. E. (2007). Influences on bicycle use.*Transportation*, *34*(4), 453-470.